Genetic variations in the β-globin cluster: influence on levels of fetal hemoglobin

Raquel Alexandra Jesus Moreira dos Santos

Mestrado em Genética Forense

Departamento de Biologia

2019

Orientador

Maria João Prata, PhD, Faculdade de Ciências da Universidade do Porto (FCUP); Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP); Instituto de Investigação e Inovação em Saúde (i3S).

Coorientadores

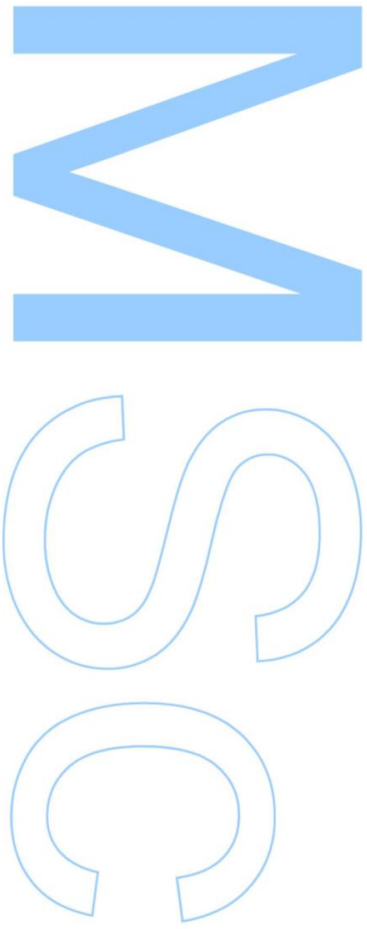
Licínio Manuel Manco, PhD, Faculdade de Ciências e Tecnologia da Universidade de Coimbra (FCTUC); Centro de Investigação em Antropologia e Saúde (CIAS); Centro Hospitalar e Universitário de Coimbra (CHUC).

Verónica Gomes, PhD, Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP); Instituto de Investigação e Inovação em Saúde (i3S).

All corrections determined by the jury, and only those, were incorporated.

The President of the Jury,

Porto, _____/_____/_____



Agradecimentos

A realização deste projeto contou com o auxílio e colaboração de várias pessoas, às quais quero deixar o meu agradecimento.

Ao Grupo de Genética Populacional e Evolução, do Instituto de Investigação e Inovação em Saúde (i3S) pela amabilidade com que fui recebida e por me terem feito sentir em casa ao longo de todo este ano. Um especial agradecimento ao Doutor António Amorim por me ter concedido a chance de poder realizar este projeto de mestrado neste grupo.

À Professora Doutora Maria João Prata pela admirável orientação que me proporcionou ao longo de toda esta etapa, pelo total apoio, pela partilha do saber, pela dedicação, exigência, opiniões e críticas que me motivaram a ser melhor; pela paciência e pela disponibilidade. Aqui lhe deixo a minha gratidão.

À Doutora Verónica Gomes, pela coorientação deste projeto, pelo auxílio e acompanhamento em toda a parte laboratorial desta investigação. Agradeço também todos os conhecimentos que me transmitiu, sempre acompanhados da sua inconfundível simpatia.

Ao Doutor Licínio Manco, por todas as valiosas contribuições que me prestou, também na qualidade de coorientador, por me ter recebido e colaborado na realização deste trabalho. O seu apoio foi indispensável e, pela paciência e disponibilidade, estou-lhe muito grata.

À Catarina Gomes, colega de mestrado que integrou o mesmo grupo de investigação, pelo apoio, companheirismo, boa disposição e motivação que me prestou ao longo de todo o ano e que de muito me valeu em momentos menos fáceis.

Às colegas de Biologia que, embora estejam neste momento a percorrer os seus caminhos, nunca deixaram de estar do meu lado. Agradeço especialmente à Ana Santos que me apoiou e transmitiu conhecimentos valiosos e indispensáveis para a elaboração deste trabalho.

Aos meus pais que me encheram de confiança, que sempre me apoiaram e que me fazem ver o lado positivo de tudo. Nunca duvidaram de mim e sempre me fizeram acreditar que tudo é possível. São os meus pilares, a minha inspiração e aqueles a quem devo tudo aquilo que sou hoje. Sem eles, nada disto seria possível e é graças a

eles que a concretização dos meus sonhos se tornou uma realidade. Não existem palavras suficientes para exprimir a minha eterna gratidão.

Ao meu namorado, por estar sempre do meu lado, pela paciência, pela calma que me transmitiu, pela compreensão e delicadeza, pela bondade que demonstrou em todos aqueles dias e noites que sacrifiquei em prol deste projeto. Agradeço toda a preocupação e toda a força que me ajudou a ultrapassar todos os obstáculos.

A todos, o meu imenso e sincero, obrigada.

Abstract

Hemoglobin (Hb) disorders are the most common inherited blood disorders in the world. β-hemoglobinopathies, such as sickle cell disease (SCD) and β-thalassemia result from mutations in the *HBB* gene and it is known that their clinical course is ameliorated by elevated levels of HbF. This connection has prompted much interest in formulating therapeutic approaches that stimulate the production of this benign form of hemoglobin.

Recent genome wide association studies (GWAS) have identified a number of SNPs associated with an elevation on HbF levels in the general population as well as in patients with SCD and β-thalassemia.

BGLT3 is a gene that codes for an erythroid-specific long non-coding RNA (lncRNA). Its transcription as well as the transcript itself have been implicated in the regulation of γ-globin region expression, however the relationship between genetic variations at this gene and HbF levels has been not yet investigated

In order to understand the influence of this gene in HbF levels, in this study *BGLT3* was interrogated in a sample of 71 Portuguese β-thalassemia carriers. Linear regression analysis led into the discovery of a SNP (rs7924684) whose effect in HbF levels was higher than ever reported for any of the few other candidate variations implied in HbF levels.

Besides, taking advantage of a recently developed Multiplex SNaPshot® system for haplotyping the β-cluster that included six SNPs located throughout the β-globin cluster, three additional SNPs were also found to be strongly correlated with HbF levels, namely rs7482144, located in the promoter of *HBG2*, rs10128556 in the *HBBP1* pseudogene and rs968857 in the intergenic region between *HBBP1* and *HBD.*

Haplotypic analysis further allowed to identify two seven loci-defined haplotypes showing very elevated correlation with the levels of HbF.

In general, our data strengthens the evidence on the major role that sequence variation throughout the entire β-cluster plays in HbF variability. However, more important than the specific alleles present at each position, is the extended haplotypic combination of alleles at different SNPs, producing long-range sequences, that only in its whole modulate levels of HbF.

Given that the four SNPs found associated with HbF were located in non-coding regions of the β-globin cluster (promoter of *HBG2*, *BGLT3*, *HBBP1* and the intergenic region between *HBBP1* and *HBD*), further studies are needed to understand how these four regions interact to affect variability in levels of HbF.

Keywords: β-thalassemia; Fetal Hemoglobin (HbF); *BGLT3* gene; *HBBP1* pseudogene; *HBB* cluster; Single Nucleotide Polymorphisms (SNPs); SNaPshot® Multiplex; Haplotypes.

Resumo

Defeitos na hemoglobina são responsáveis pelo grupo de doenças hereditárias do sangue mais comuns no mundo. Sabe-se que as hemoglobinopatias resultantes de mutações no gene *HBB*, como a anemia falciforme e as β-talassemias, apresentam quadros clínicos mais favoráveis em doentes com níveis elevados de hemoglobina fetal (HbF). Esta relação despertou muito interesse no desenvolvimento de abordagens terapêuticas para as β-hemoglobinopatias que estimulem a produção desta forma de hemoglobina benigna.

Estudos de associação *Genome Wide* resultaram na identificação de um pequeno número de SNPs associados a elevados níveis de HbF tanto na população geral, como em pacientes com anemia falciforme e β-talassemia.

O gene *BGLT3* codifica para um *long non-coding RNA* (lncRNA) eritroide-específico. A sua transcrição, bem como o transcrito em si, foram implicados na regulação da expressão da região da γ-globina, contudo a relação entre variações genéticas neste gene e níveis de hemoglobina fetal ainda não foi investigada.

De modo a compreender a influência de variações em *BGLT3* nos níveis de hemoglobina fetal, neste estudo, o gene foi investigado numa amostra de 71 portugueses portadores de β-talassemia. Análise de regressão linear permitiu encontrar um SNP em *BGLT3* (rs7924684) cujo efeito nos níveis de HbF era muito superior a qualquer um dos descritos para as poucas variações genéticas já implicadas em níveis de HbF.

Por outro lado, tirando partido de um sistema *SNaPshot® Multiplex* desenvolvido para haplotipar o cluster das β-globinas, que incluía seis SNPs dispersos pelo cluster, detetaram-se três SNPs adicionais também fortemente relacionados com níveis de HbF, nomeadamente rs7482144, localizado no promotor de *HBG2*, rs10128556, no pseudogene *HBBP1* e rs968857, na região intergénica entre *HBBP1* e o gene *HBD.*

A análise haplotípica permitiu ainda identificar dois haplótipos definidos por sete loci que demonstraram uma correlação muito elevada com os níveis de hemoglobina fetal.

Em geral, os nossos dados fortalecem a noção de que variações de sequência ao longo de todo o cluster β desempenham um papel muito importante na variabilidade quanto aos níveis de HbF.

Dado que os quatro SNPs que revelaram associações significativas se situam em regiões não codificantes do cluster β (promotor do *HBG2*, *BGLT3, HBBP1* e a região

intergénica entre *HBBP1* e *HBD*), mais estudos são necessários para compreender como estas quatro regiões interagem para afetar a variabilidade quanto aos níveis de hemoglobina fetal.

Palavras-chave: β-talassemia; Hemoglobina Fetal; *BGLT3*; Cluster *HBB*; *Single Nucleotide Polymorphisms* (SNPs); SNaPshot® Multiplex; Haplótipos.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A	Adenine
α	Alpha
$\alpha_2\varepsilon_2$	Gower Hemoglobin 2
AI	Arab-Indian
AML	Acute Myeloid Leukemia
AP-1	Activator Protein 1
β	Beta
BAN	Bantu
BCL11A	B-Cell Lymphoma/Leukemia 11A
BEN	Benin
BGLT3	Beta Globin Locus Transcript 3
bp	Base Pair
BP-1	Beta Protein 1
C	Citosine
CAM	Cameroon
CAR	Central African Republic
CO_2	Carbon Dioxide
◦C	Degree Celcius
δ	Delta
DNA	Deoxyribonucleic Acid
ε	Epsilon
G	Guanine
γ	Gamma
GWAS	Genome Wide Association Study
Hb	Hemoglobin
HbA1	Adult Hemoglobin - major form
HBA1 (α1)	Alpha 1 Globin Gene
HbA2	Adult Hemoglobin - minor form
HBA2 (α2)	Alpha 2 Globin Gene
HBB (β)	Beta Globin Gene
HBB*S	Sickle Cell Allele
HBBP1 (ψβ)	Beta Globin Pseudogene
HBD (δ)	Delta Globin Gene
HBE (ε)	Epsilon Globin Gene
HbF	Fetal Hemoglobin
HBG1 (Aγ)	Gamma A Globin Gene
HBG2 (Gγ)	Gamma G Globin Gene
HPFH	Hereditary Persistance of Fetal Hemoglobin
HPLC	High Performance Liquid Chromatography
HS	Hipersensitive Site

HSCT	Hematopoietic Stem Cell Transplant
HW	Hardy-Weinberg
∞	Infinite
KLF1	Kruppel-like Factor 1
LCR	Locus Control Region
LD	Linkage Disequilibrium
LncRNA	Long Non-coding RNA
log	Logarithm
LRF	Lymphoma/Leukemia-related Factor
µL	Microliter
µM	Micromolas
min	Minute
NF-E2	Nuclear Factor Erythroid 2
NuRD	Nucleossome Remodeling Deacetylase
O_2	Oxygen
ψ	Psi
%	Percentage
PCR	Polymerase Chain Reaction
RBC	Red Blood Cell
®	Registered
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
rs	Reference SNP
SBE	Single Base Extension
SCD	Sickle Cell Disease
SCL/TAL1	Stem cell leukemia/T-cell acute lymphoblastic leukemia 1
sec	Second
SEN	Senegal
SNP	Single Nucleotide Polymorphism
T	Timine
™	Trade Mark
V	Volt
v.	Version
ζ	Zeta
ζ2γ2	Portland Hemoglobin
ζ2ε2	Gower Hemoglobin 1
ZBTB7A	Zinc finger and BTB domain-containing protein 7A

1. Introduction

1.1. Hemoglobin

Hemoglobins (Hbs), the proteins that give the recognized red color to the blood, play a pivotal role in the human physiology, and consequently in human health. Their major function is to carry oxygen (O_2) from the lungs to peripheral organs (e.g. muscles), where O_2 is used for aerobic metabolism, and also help carry the product of the aerobic metabolism - carbon dioxide (CO_2) – from those organs to the lungs. They transport these gases within red blood cells (RBCs) (Honig *et al.* 1986; Sherwood, 2015; Philipsen *et al.* 2018; Wienert *et al*. 2018).

The Hb molecule consists of four polypeptide chains, two α-like and two β-like globins (Figure 1). Nascent globin chains rapidly incorporate a heme group, which stabilizes their native folding into Hb subunits (Thom *et al.* 2013). Since each α and β subunit forms a dimer, Hb is often referred to be a dimer of αβ dimers. During development, the type of α and β chains synthetized changes, giving rise to different tetrameric combinations, and thus to distinct forms of Hb. Accordingly, depending on the developmental stage, embryonic, fetal or adult Hbs are produced (Figure 2), all with the typical heme groups bounded, since the later are necessary for guaranteeing the molecule's oxygen-carrying capacity (Adelvand *et al.* 2017; Sankaran *et al.* 2013). At the embryonic stage, the proteins produced are known as Hb Gower 1 ($\zeta_2\varepsilon_2$), Hb Gower 2 ($\alpha_2\varepsilon_2$) and Hb Portland ($\zeta_2\gamma_2$), appearing in this order. At the fetal stage, the prevailing Hb is HbF ($\alpha_2\gamma_2$), while at the adult stage, the RBCs produce HbA1 ($\alpha_2\beta_2$) and HbA2 ($\alpha_2\delta_2$) (Marengo-Rowe, 2006; Adelvand *et al.* 2017). The two kinds of polypeptide chains, the α and the β globins are similar in length but differ in amino acid sequences. The different globin chains are controlled by two clusters of genes: one located on chromosome 16 that contains seven loci including two codifying for α (*HBA1* and *HBA2*), and other, located on chromosome 11 that comprehends one pseudogene and five genes codifying β-like globins (*HBE*, *HBG2, HBG1, HBB* and *HBD*) (Weatherall *et al.* 2006).

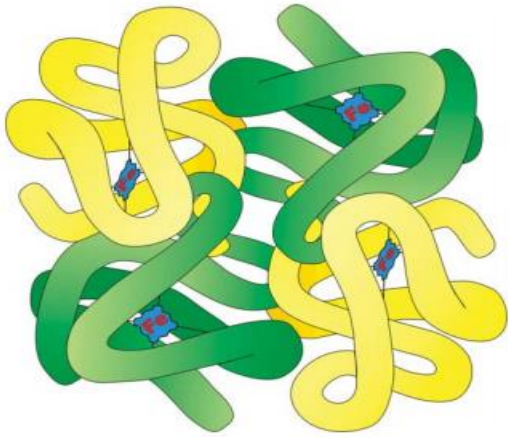


Figure 1 – Schematic diagram showing the basic structure of a single hemoglobin molecule, including two α-globin chains (green), two β-globin chains (yellow), each one containing a heme-iron complex (blue) (Thomas *et al.,* 2012)

1.1.1 Hemoglobin Switch

In humans, two developmental switches occur in the expression of genes from the β-globin cluster: an ε-to-γ globin switch and a γ-to-β switch (Figure 2). Within the first trimester of pregnancy, ε-globin gene is dominantly expressed. As its expression is silenced, the γ-globin genes are upregulated. Then, shortly after birth, the second switch is completed when γ-globin gene expression is virtually silenced and the adult β-globin becomes the major β-like globin. δ-globin is also produced in adulthood, however, it represents only a small amount of the total β-like globins (Martyn *et al.* 2018; Mozeleski *et al.* 2018; Wienert *et al.* 2018). The sequential activation and silencing of globin genes are fine-tuned controlled (Shang *et al.* 2017).

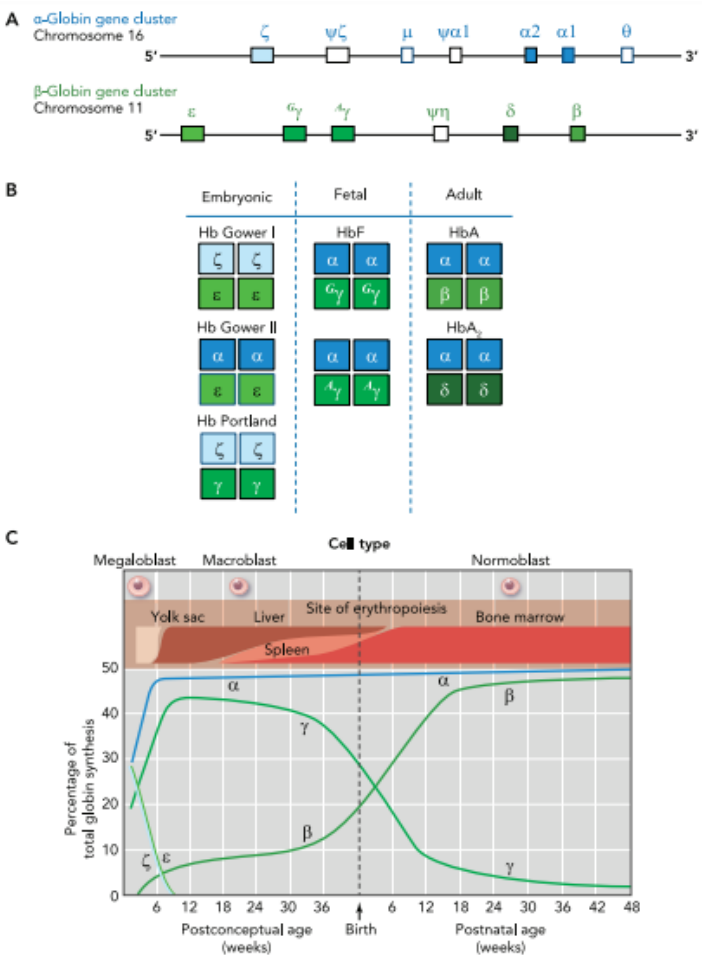


Figure 2 – The expression of α- and β-type globin genes. **A**: diagram of the human α- and β-globin gene clusters. **B**: set of distinct embryonic, fetal and adult Hb and their different tetrameric combinations. **C**: Hemoglobin switch - The ε-globin is expressed during the embryonic stage and replaced by γ-globin during fetal life. Around time of birth, a γ-to-β-globin switch occurs and the β-globin is predominantly expressed in the adult life. The adult δ-globin gene is poorly expressed (Storz, 2016; Cavazzana *et al.*, 2017).

Recent studies have identified several players in the regulation of the globin chain switching in the β-cluster. The process is controlled by a number of transcription factors that interact directly with genes from the cluster, or indirectly through connections with other transcription factors to influence gene expression (Habara *et al.* 2017). Among the already known factors are included the erythroid Kruppel-like factor (*KLF1*); the stage selector protein *MYB*; and the B-cell lymphoma/leukemia 11A (*BCL11A)* and lymphoma/leukemia-related factor (*LRF)*, the two major repressors of γ-globin regulation (Borg *et al.* 2010; Stadhouders *et al.* 2014; Martyn *et al.* 2018; ; Das *et al.* 2019). These transcription factors and the mechanisms through which they control the switching process will be further detailed in topic 1.1.4.1.

1.1.2 β-globin Gene Clusters

The β-globin gene cluster is located in an AT rich region of the short arm of the human chromosome 11, band p15.5, covering roughly 702kb of DNA and consists of five coding genes that are specifically expressed in erythroid cells: the embryonic-specific ε (*HBE1*), the fetal-predominant Gγ and Aγ (*HBG2* and *HBG1*), and the late expressed δ and β (*HBD* and *HBB*) globin genes. The intergenic region between the Aγ- and δ-globin genes contains a pseudogene (*HBBP1*) and a noncoding gene (*BGLT3*) (Figure 3; Brittain, 2002; Wienert *et al.* 2018; Huang *et al.*, 2017).

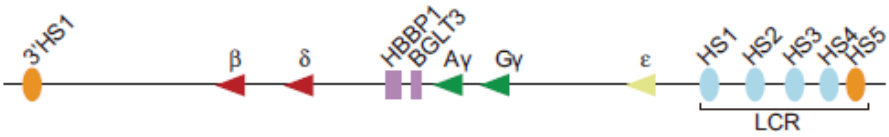


Figure 3 – Diagram of the human β-globin locus (Huang *et al.*, 2017)

The expression of all the β-globin genes requires a strong enhancer, the locus control region (LCR). The LCR is a 16-kb-long cis element and regulatory region located upstream of the ε-globin gene that contains powerful chromatin opening capacity and five DNA enhancer elements that are DNase hypersensitive sites (HSs), which are necessary for high-level globin gene expression. In order to regulate the expression of the β-globin genes, the LCR interacts directly with their promoters by looping in a dynamic manner to the fetal γ-globin and adult β-globin promoters in fetal and adult erythroblasts, respectively (Huang *et al.* 2017; Cavazzana *et al.* 2017; Adelvand *et al.* 2017).

1.1.3 Hemoglobin Genetic Disorders

Hb disorders are the most common inherited blood disorders in the world and account for approximately 3.4% of deaths in children under 5 years old (Modell *et al.* 2008). Since these genetic disorders are due to globin gene mutations, they are broadly subdivided into two general categories, according to the molecular defect: those in which gene mutations impair globin protein subunit production (thalassemias), by greatly reducing it

or ceasing it completely, and those in which the mutation produces a structural defect in one of the globin subunits (Hb variants). The latter class is mostly composed of missense mutations that cause single amino acid substitutions in the globin protein, resulting in a "variant" Hb tetramer (Forget *et al.* 2013; Thom *et al.* 2013).

β-hemoglobinopathies, such as sickle cell disease (SCD) and β-thalassemia which are the most common monogenic disorders worldwide, result from mutations in the *HBB* gene (Martyn *et al.* 2018; Cavazzana *et al.* 2017). In general, β-hemoglobinopathies follow a recessive mode of inheritance, albeit rare cases of dominant β-thalassemias exist. Despite falling in the designation of Mendelian single gene disorders, individuals with the same disease or even the same genotype can display extreme clinical heterogeneity (Weatherall *et al.* 2006; Steinberg *et al.* 2001; Galanello *et al.* 2010; Liu *et al.* 2016).

SCD is caused by an A-to-T point mutation in the sixth codon of the β-globin gene that replaces glutamic acid with valine in the β-globin polypeptide. This results in the polymerization of the defective β-globin (β^S) in its deoxygenated state, making the normal RBC become inflexible and hook/sickle-like shaped, which in turn causes the obstruction of microvessels, impairing the function of many organs and contributing to increased morbidity and early mortality (Cavazzana *et al.* 2017; Marengo-Rowe, 2006; Liu *et al.* 2016).

β-thalassemias are caused by β-globin gene mutations that reduce or abrogate the production of β-globin chains (β^+ and β^0 genotypes, respectively). More than 200 different β-thalassemia mutations have been reported up to now (Hardison *et al.* 2002). Depending on its clinical severity, the disease can be categorized into β-thalassemia major (the most severe phenotypic manifestation of the disease), β-thalassemia minor (often clinically silent) and β-thalassemia intermedia (with phenotypic manifestations that lie between the other two forms). Patients with β-thalassemia minor and intermedia can survive without regular blood transfusions, however those affected by β-thalassemia major are transfusion-dependent, as they suffer from many complications, such as severe anemia (Cavazzana *et al.* 2017; Galanello *et al.* 1998).

The clinical course of β-hemoglobinopathies is ameliorated by elevated levels of HbF which reduces globin chain imbalance in β-thalassemias and exerts a potent anti-sickling effect in SCD (Antoniani *et al.* 2018).

Currently, the major treatments for these disorders involve symptomatic care and blood transfusions. Hydroxyurea is a small molecule that boosts the production of HbF in many patients and is now commonly used in the treatment of patients with SCD and some with

β-thalassemia. Bone marrow transplantation from a sibling, matched unrelated donor or, more recently, haploidentical transplantation is also included in the treatment options, but the availability of donors is limited and the procedures carry substantial risks, such as iron overload, transmission of infectious diseases and the possibility of developing RBC alloantibodies. Allogenic hematopoietic stem cell transplant (HSCT) can be curative for SCD but only 18% of SCD patients have matched sibling donors. Receiving a graft from an unmatched donor is associated with a higher risk for mortality and morbidity (Steinberg *et al.* 2014; Steinberg *et al.* 2012; Ferrone, 2016; Steinberg *et al.* 2010; Voskaridou *et al.* 2010; Wong *et al.* 2014). There are many challenges in the implementation of these treatments, specially the limitation in their use, particularly in the developing world (Roseff, 2009; Michlitsch and Walters, 2008; Persons, 2009; Sankaran *et al.* 2010b; Wienert *et al.* 2018).

Giving the variety of clinical observations that demonstrate the capacity of HbF in ameliorating the severity of β-hemoglobinopathies, there has been an interest in formulating therapeutic approaches that stimulate the production of this form of hemoglobin (Sankaran *et al.* 2013). Recently, hemoglobin research has regained momentum due to the development of genome editing tools, such as Clustered, Regularly Interspaced, Short Palindromic Repeat (CRISPR) – associated 9 (Cas9) (CRISPR-Cas9). The ability to introduce genome modifications means that a routine gene therapy cure for these diseases can now be a real possibility (Wienert *et al.* 2018).

Therapeutic genome editing strategies targeting the HbF repressors directly are already under consideration. By targeting repressors, such as BCL11A or LRF, using genome editing to delete them or reduce their expression it may be possible to elevate HbF levels. However, transcription factors usually have other roles, regulating multiple target genes, therefore some unwanted side effects can be encountered. Another way of obtaining higher HbF levels could be through avoiding the action of cis-regulatory repression elements, specifically at the γ-globin locus, by mutating their binding sites, for example, mimicking HPFH by introducing know HPFH mutations, which is being attempted in several laboratories, at current time (Wienert *et al.* 2018; Antoniani *et al.* 2018; Traxler *et al.* 2016).

1.1.3.1. β-thalassemia in Portugal

β-thalassemia is the most common recessive inherited disorder in Mediterranean populations (Weatherall, 1986), including in the Portuguese, in which it occurs rather frequently (Coutinho-Gomes *et al.* 1988).

Pioneer epidemiological studies of β-thalassemia in Portugal revealed it was present at 0.45% frequency, though unevenly distributed across the country with higher prevalence in the central and southern regions (Martins *et al.* 1993). Later, the genetic characterization of β-globin mutations in Portuguese thalassemic individuals was initially focused mainly in the areas with the highest incidence of the disease which fueled the assumption of a low or even null incidence of the disease in the North (Coutinho-Gomes *et al.* 1988; Lavinha *et al.* 1992; Faustino *et al.* 1992; Tamagnini *et al.* 1983; Martins *et al.* 1993). However, another study, specifically addressing β-thalassemia in northern Portugal, came to disprove not only the absence of the disease in the region, but also the previous belief in a great homogeneity of β-thalassemia mutations in Portugal, since an almost absent mutation in the central and southern areas (CD6(-A)) was found in 40% of the affected families from the North of the country (Cabeda *et al.* 1999).

According to a recent review on the mutational spectrum of β-thalassemia in Portugal, the six most frequent mutations are CD 39, IVSI-1, IVSI-6, IVSI-110, CD15 (TGG-TGA) and CD 6 (-A) (Almeida, 2015). In the southern/central Portuguese prevail the C 39, IVSI-1 and IVSI-110, while in northern Portuguese CD 6 (-A) dominates, although CD 39 and IVSI-1 are also well represented.

A note on the IVSI-6 mutation, which was originally described in a group of thalassemia intermedia patients from a small area in northern Portugal (Tamagnini *et al.* 1983). Patients homozygous for this mutation presented with a remarkably mild clinical course, as evidenced by their age and the absence of transfusion dependence, despite extremely low levels of HbF (<20%). Since a combination of a mild clinical course (without any ameliorating factors) and low HbF levels had not been previously described, at the time, it was proposed that the variant should be known as β^+ thalassemia Portuguese type (Tamagnini *et al.* 1983). Soon after, however, it was shown that IVSI-6 was rather common across peri Mediterranean regions (Cao *et al.* 1989), and was not restricted to the North of Portugal, since in the South/Center it also occurred, even at lower frequency in comparison with the North (Cao *et al.* 1989; Almeida, 2015).

1.1.4 Genetic Modifiers of Fetal Hemoglobin (HbF)

1.1.4.1 First Insights on Quantitative Trait Loci

HbF is the main hemoglobin throughout the fetal life. At birth, it accounts for approximately 80% of total hemoglobin, while in adults it is only residually produced,

representing less than 1-2% of total hemoglobin (Mandal *et al.* 2019; Martyn *et al.,* 2018), although this level varies considerably in healthy subjects.

A greater than usual elevation of HbF is the hallmark of Hereditary Persistence of Fetal Hemoglobin (HPFH), an inherited condition characterized by persistent high (>2%) levels of HbF in adulthood. Most commonly, people with HPFH are asymptomatic and appear to be healthy, indicating that high levels of HbF are benign. The condition is characterized by elevated production of γ-globin in adulthood, consequently resulting in high HbF levels. According to the molecular defects that lead to HPFH, the condition is classified as one of two types: (1) deletional HPFH or (2) non-deletional HPFH. Deletional HPFH is due to large deletions within the β-globin gene cluster (3' to the γ-globin genes), including partial or full deletion of *HBB* or *HBD*, and thus the high HbF levels are associated with variable compensation for the partial or total lack of *HBD* and/or *HBB* gene expression. Non-deletional HPFH is caused by single base substitutions, or small deletions, in one of the two γ-globin genes promoters (Patrinos *et al.* 2004; Giardine *et al.* 2007; Giannopoulou *et al.* 2012; Orkin *et al.* 2019). The detection that such alterations fall into two distinct clusters located approximately 115 and 200bp upstream of the γ-globin genes transcriptional start sites, suggests that most likely those HPFH-associated mutations disrupt the binding of transcriptional repressors that normally silence the γ-globin genes (Martyn *et al.* 2018) leading to the reactivation of fetal globin gene transcription.

Despite the progresses in understanding HPFH, actually mendelian forms of HPFH are rare and do not explain common HbF variability in healthy adults.

While the persistence of high levels of HbF has no clinical consequence in healthy individuals, HbF is assuming greater and greater importance because it is now widely demonstrated that increased HbF levels alleviate the clinical severity of the two major β-hemoglobinopathies, namely SCD and β-thalassemia (Wang *et al.* 2018).

For that reason, strong efforts have been made in the last few years to decipher the genetic architecture underlying variability in HbF levels.

Besides the mutations already known yielding a benign HPFH phenotype, which in general are not relevant to HbF levels neither in healthy subjects nor in patients affected with hemoglobinopathies, there are a number of common variants, i.e., with allele frequency above 1% that act together in a polygenic manner, which have been identified as amounting to the overall variability in HbF phenotypes in the general population as well in patients with SCD and β-thalassemia (Menzel and Thein, 2018). The discovery was the fruit of recent Genome wide association studies (GWAS), conducted in both

healthy subjects and in patients with β-hemoglobinopathies of European, African or Asian descent, that allowed the identification of several SNP variants associated to HbF levels.

The β-globin locus has been at the forefront of studies on the dynamics and mechanisms of chromatin looping and gene regulation. As previously mentioned, in human erythroid cells there is a powerful distal enhancer called LCR that gains proximity with the β-globin genes in a developmentally dynamic way (Carter *et al.* 2002; Tolhuis *et al.* 2002). In primitive murine blood cells, the LCR loops to the embryonic type β-globin gene promoters, resulting in the dominant expression of ε-globin expression. In adult erythroid cells, the LCR exclusively contacts the adult β-globin genes (*HBB* and *HBD*). Also, in the fetal stage of the hemoglobin switch (from the beginning of fetal liver erythropoiesis until birth), the LCR loops to the γ-globin genes, enabling their upregulation. (Palstra *et al.* 2003).

Several transcription factors have been implicated in loop formation at the β-globin locus (Figure 4). The appearance of these specific repressor proteins during development renders the promoters of the embryonic/fetal globin genes inaccessible for activation, shifting the LCR-promoter interactions to the adult globin genes (Philipsen, 2018).

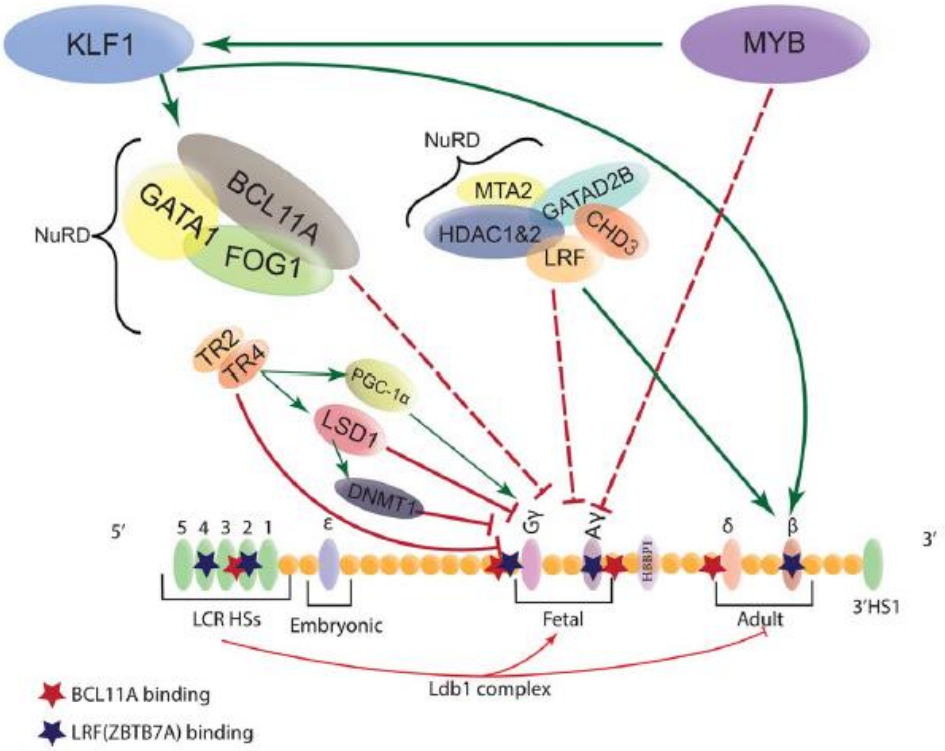


Figure 4 – Cis- and trans-acting effectors of gene expression within the β-globin gene cluster (Habara *et al.* 2017)

In the earliest studies performed, in the decades following the molecular cloning of the globin genes, a number of transcription factors implicated in the regulation of β-globin genes were identified, namely GATA1, KLF1 and SCL/TAL1 (Sankaran *et al.* 2013). However, at the time, none of these factors seemed to be specific regulators of the hemoglobin switch. It was only nearly three decades later that this process began to be fully understood, with the identification of other transcriptional factors, such as LRF, encoded by ZBTB7A and BCL11A, now known to be major repressors of HbF (Sankaran *et al.* 2013).

The Leukemia/lymphoma-related factor (LRF), is a zinc finger transcription factor demonstrated to be an important repressor of γ-globin gene expression. It affects HbF gene silencing through its binding in and about the *HBB* gene cluster (Masuda *et al.* 2016; Habara *et al.* 2017).

The B cell CLL/lymphoma 11A (BCL11A), also a zinc finger protein, has been intensively studied (Sankaran *et al.* 2015; Zhou *et al.* 2010; Sankaran *et al.* 2008; Lettre *et al.* 2016) and reported as a major repressor of HbF production. It interacts with other silencing factors, such as GATA1 and the NuRD complex, at specific sequences in the human β-globin gene cluster to silence γ-globin gene expression by promoting long-range physical

interactions between the LCR and the *HBB* promoter, at the expense of the γ-globin gene promoter (Xu *et al.* 2010; Sankaran *et al.* 2011; Sankaran *et al.* 2008).

Some of the transcription factors involved in HbF regulation are cis-acting elements, binding to the specific site they regulate. Some examples are GATA1, NF-E2 and EKLF (also known as KLF1). The GATA binding protein 1 (GATA-1) is involved in the human adult hemoglobin stage by mediating a loop between the DNase I Hypersensitive site 1 (HS1) in the LCR and *HBB* (Adelvand *et al.* 2017). The nuclear factor erythroid 2, known as NF-E2 interacts with the promoter of *HBB* at the human adult stage in order to help establish a connection with the LCR. Finally, the erythroid-Krüppel-like factor (EKLF), or Krüppel-like factor 1 (KLF1) participates in the γ-to-β hemoglobin switch through direct binding to the promoters of the correspondent genes (Adelvand *et al.* 2017; Cho *et al.* 2008; Catani *et al.* 2002; Zhou *et al.* 2006; Alhashem *et al.* 2011).

Other transcription factors (or even the same ones that are cis-acting elements at one point) behave as trans-acting elements. These elements code for a protein, microRNA or other diffusible molecule that will be used in the regulation of another target gene as, for instance, KLF, NF-E2, BP-1, BCL11A and the NuRD complex (Adelvand *et al.* 2017).

KLF1 belongs to a family of zinc finger (Znf) proteins that are involved in human embryonic erythropoiesis and the regulation of the β-globin gene expression and is controlled by the stage selector protein MYB. KLF1 is involved in the mechanism that silences γ-globin gene expression by directly regulating (in this case, activating) genes, such as BCL11A (by occupying promoter and activating its expression) and MYB (Adelvand *et al.* 2017; Tallack *et al.* 2013; Borg *et al.* 2010; Zhou *et al.* 2010).

NF-E2 is a heterodimeric transcription factor that pertains to a group of proteins called the basic-leucine zipper (bZIP) proteins. It is a specific erythroid activator that couples with GATA-1 in order to create a chromatin loop between the γ-globin gene and the HS2 region of the LCR (Andrews, 1998; Woon Kim *et al.* 2011).

The Beta Protein 1 (BP-1) is thought to be an oncogenic molecular marker associated with acute myeloid leukemia (AML). It negatively controls the expression level of the β-globin gene, during the embryonic and fetal erythropoiesis, through the blocking of GATA-1 and KLF1 (Mpollo *et al.* 2006).

The Nucleosome Remodeling Deacetylase (NuRD) Complex is a specific silencer of γ-globin gene expression and it inhibits γ-globin gene transcription during the developmental stage through an epigenetic mechanism that recruits seven polypeptides

(Harju-Baker *et al.* 2008). It consists of several components that control/regulate various other factors. For example, one of its components is the Mi2β protein that directly binds and positively regulates γ-globin gene silencers, such as KLF1 and BCL11A (Adelvand *et al.* 2017; Amaya *et al.* 2013).

Up to now, among the most well-established genetic factors associated with HbF levels are included three SNPs, specifically lying in *HBG2,* BCL11A and *HMIP,* concerning which Pereira *et al.* succeeded to replicate that they were strongly correlated with increased HbF levels in Portuguese patients with β-hemoglobinopathies (Pereira *et al.* 2015).

1.1.4.2 *BGLT3* Gene

Very recently, a new locus was uncovered that displays enhancer-like features increasing the production of HbF. It is *BGLT3* (Beta Globin Locus Transcript 3), a gene lying downstream of the duplicated γ-globin genes and upstream of the adult δ-globin and β-globin gene (Figure 4) that codes for an erythroid-specific long non coding RNA (lncRNA). LncRNAs are increasingly being appreciated as participants in the regulation of important cellular processes, including transcription. Since most of them are highly cell-type specific (for example, hundreds are known to be expressed specifically during erythropoiesis) they are assumed to play an important role in the generation of diverse cell types and in cell-specific functions (Ivaldi *et al.* 2018).

Concerning *BGLT3*, its transcription was proven to be consistently associated with γ-globin transcription in erythroid cells *in vivo* (Ivaldi *et al.* 2018). Remarkably, the locus locates in the intergenic region between the γ-globin and δ-globin genes, a region that since long was known to influence, though obscurely, the γ-globin expression, once early work had revealed that a number of mutations occurring there, especially large deletions, led to HPFH because the fetal globin genes continued to be expressed throughout life. For long, it was thought that the deletions either removed repressive elements or juxtaposed downstream enhancers close to the fetal globin genes, but still the results from transgenic mice or gene editing in human cells did not always recapitulate the human phenotypes (Blobel and Crossley, 2019). Evidence exists now that this intergenic region has additional important features, most notably the two genes that produce noncoding RNA, *BGLT3* and the pseudogene *HBBP1* (presented in the next point).

In 2011, Keifer *et al,* investigating NLI (Ldb1 homolog) complex occupancy and chromatin conformation of the β-globin locus in human erythroid cells, found for the first

time robust NLI complex occupancy at a site downstream of the γ-globin genes within sequences of *BGLT3*, affording evidence that in those cells' proximity between the *BGLT3*/ γ-globin region and the LCR was established (Keifer *et al.* 2011). This work led to conclude that γ-globin transcription or silencing was mediated through long-range LCR interactions that involved looping from the *BGLT3* sequences to *HBG1* and *HBG2*.

Later, mechanistic studies demonstrated that the *BGLT3* transcript as well as the *BGLT3* transcription itself, were both positive regulators of γ-globin region expression, although being implied in distinct looping interactions (Ivaldi *et al.* 2018). In fact, while the transcription through the *BGLT3* locus was required for looping to the γ-globin genes, the *BGLT3* transcript was not, but instead it interacted with the mediator of RNA polymerase II transcription (Mediator) complex, which is a large complex with modular organization, generally required for transcription by RNA polymerase II regulating various steps of this process (Keifer *et al.* 2011; Ivaldi *et al.* 2018).

In short, it was proposed that *BGLT3* operates via multiple mechanisms, including maintaining an enhancer-like state, looping to the γ-globin genes and recruiting co-regulators, possibly through the *BGLT3* transcript itself (Blobel and Crossley*,* 2019).

Understanding the influence of *BGLT3* in HbF levels is of utmost relevance because, as previously mentioned, elevated HbF level alleviate the clinical severity of β-hemoglobinopathies, prompting pharmacological and genomic approaches for therapeutic HbF reactivation or induction (Wang *et al.* 2018).

In a broadest sense, the increased knowledge on the genetic factors accounting for HbF, a strong modifier of sickle cell disease severity, for instance, can open new perspectives to investigate the cause of death in forensic cases that involve postmortem analyses (Hammer, 2006).

1.1.4.3 *HBBP1* pseudogene

In 2010, two independent GWASs led to identify two SNPs, rs10128556 and rs2071348, both residing within *HBBP1*, associated with elevated HbF levels (Galarneau *et al.* 2010; Nuinoon *et al.* 2010). In the investigation of Galarneau *et al,* the previously well-established relationship between rs7482144-XmnI and HbF levels was clearly replicated, but in addition it was discovered that rs10128556, located downstream of *HBG1*, more precisely in the second intron of *HBBP1*, was still more strongly associated with HbF levels than rs7482144-XmnI, in fact by two orders of magnitude.

In turn, the study conducted by Nuinoon *et al,* showed that the SNPs revealing the strongest association with disease severity in Thai β^0-thalassemia/HbE patients were located in the β-globin gene cluster. Out of the ten SNPs there identified, the one showing the most significant association was rs2071348, also located in the second intron of *HBBP1*.

Since rs10128566 (11:5242453) and rs2071348 (11:5242916) are extremely nearby from each other (less than 500bp separate them), it's not surprisingly that they are in total or quasi-total LD in populations, rendering extremely difficult to discriminate which of them exert higher influence in HbF levels/disease severity or contrarily if they act synchronically.

Even so, the two studies considered as a whole strongly suggested the presence of developmental regulatory elements in the region encompassing the pseudogene, and very likely they could be implied in the dynamics of chromatin structure in the entire β-globin gene cluster. As a matter of fact, previous lines of evidence had already indicated that changes in chromatin architecture should be involved in the control of gene expression in the cluster and that the process needed to be tightly coordinated. All this prompted further investigations, among which is a recent study based on the comparative high throughput analysis of three-dimensional chromosomal architecture that succeeded to demonstrate at a finer scale not only distinct chromatin folding patterns at the developmentally controlled β-globin locus, but also previously unknown stage-specific chromatin contacts involving specifically the *HBBP1* region (Huang *et al*. 2017).

In the study of Huang *et al,* the *BGLT3* region, which is adjacent to the *HBBP1* region, was also analysed and shown to be involved in chromatin interactions, though residing in a sharply separated developmental stage-specific chromatin contact domain. The then performed capture-C experiments revealed that *HBBP1* and *BGLT3* demarcated not only a functional separation of fetal vs. adult gene expression, but also that the two adjacent noncoding genes functioned by different mechanisms to facilitate globin switching. *HBBP1* seemed to contribute to fetal globin repression, possibly by binding the repressor *BCL11A,* and in contrast to *BGLT3*, the *HBBP1* gene region, and not the transcript itself or the act of transcription through the gene, was required for influencing the architecture of the locus, presumably by separating the fetal or adult globin genes from the LCR enhancer at different developmental stages (Huang *et al.* 2017; Blobel *et al.* 2019).

1.2 β-globin Haplotypes

SNPs are the most abundant type of genetic variation and are estimated to occur at 1 out of every 1000 bases in the human genome. Depending on where they occur (coding or non-coding regions) they can have different consequences at the phenotypic level. As with other kind of genetic marker, different SNPs can be in Linkage Disequilibrium (LD) which is defined as the non-random association of alleles at different loci, referring thus to the fact that non-alleles can co-occur on the same haplotype more or less often than expected by chance (Crawford *et al.* 2005; Sachidanandam *et al.* 2001; Venter *et al.* 2001; Syvänen 2001; Wall *et al.* 2003).

Haplotypes are a combination of alleles at different markers that are inherited as a unit unless recombination or mutation occurs. For autosomes, each individual has two copies for a given stretch of the genome, representing the maternally and paternally transmitted haplotypes (Crawford *et al.* 2005).

Haplotyping of variants within the β-globin cluster has revealed to be very productive. The earliest analyses were crucial to illuminate the pattern of recombination within the region and to address questions from the population genetics' field or related with the evolution of β-globin genes or variants (Crawford *et al.* 2002; Vinson *et al.* 2004). Among the latter, studies on the origin of the β^S mutation represented an important topic of research. Early in those studies, it was demonstrated that the β^S mutation was found only in a few haplotypic backgrounds. According to their prevalence in different ethno-linguistic groups or geographical regions, the five main haplotypes (Table 1.) began to be referred as: the Bantu/Central African Republic (BAN/CAR) haplotype, which dominates in South-Central and Eastern Africa; the Benin (BEN), prevailing in the African Midwest; the Senegal (SEN) characteristic of Atlantic Africa; the Cameroon (CAM) commonly found within the geographical boundaries of Cameroon and at a smaller frequency in the west coast of Africa; and the Arab-Indian (AI) haplotype mainly present in the Arabian Peninsula and India. The wide geographical distribution of β^S together with the fact of being anchored in different haplotypic backgrounds, appeared to sustain the multicentric model on the origin of the β^S allele, according to which each haplotype represented an independent occurrence of the same exact mutation in a given geographic region. Albeit it was the most accepted model for decades, it co-existed with another hypothesis on the origin of the sickle mutation, positing instead its single and recent origin, a scenario that very recently gained renewed evidence based on the analysis of whole-genome-sequence data mutation (Shriner and Rotimi, 2018), meaning thus that the issue still remains debatable.

A very important observation that came with the β^S haplotyping, was that each haplotype was associated with a characteristic average level of HbF, and consequently with the severity of the clinical symptoms (Powars, 1991; Steinberg, 2009). The SEN and AI haplotypes were associated with higher levels of HbF (>15%) and a milder course of disease; BEN and CAM were related to intermediate HbF levels (5 – 15%) as well as to an intermediate clinical course; and the Bantu/CAR haplotype was related with lower levels of HbF (<5%) and more severe clinical presentation (Crawford *et al.* 2002; Leal *et al.* 2016; Hirokawa *et al.* 1995).

In fact, the establishment of such relationships would represent the first suggestion that cis-acting elements within the β-globin cluster were modulators of the HbF levels (Habara *et al.* 2017).

Currently, haplotype analysis of SCD is considered as a useful marker for disease severity, ultimately acknowledging the contribution of the haplotypic backgrounds to explain a substantial proportion of phenotypic heterogeneity in SCD (Powars, 1991; Steinberg, 2009).

Originally, *HBB* haplotypes were categorized after the examination of the patterns of restriction enzyme digestion using a limited number of restriction endonucleases, and consequently of cleavage sites, most of which were later found to be SNPs (Table 1).

Enzymes / *Haplotypes*	*XmnI (5'Gγ)*	*HindIII (Gγ)*	*HindIII (Aγ)*	*HincII (3'Ψβ)*	*HinfI (5'β)*
Senegal	+	+	−	+	+
Bantu/Central African Republic	−	+	−	−	−
Cameroon	−	+	+	+	+
Benin	−	−	−	+	−
Arab-Indian	+	+	−	+	−

(+) = Cut by a specific restriction endonuclease; (−) = Is not cut by that specific restriction endonuclease.

Table 1. Restriction endonuclease cutting patterns that represent each of the five β^S haplotypes (adapted from Bitounqui *et al.* 2015).

Any haplotype not fitting into the five standard patterns associated with the β^S mutation is conventionally referred to as atypical haplotype (Leal *et al.* 2016; Embury *et al.* 1994).

Haplotyping of β-thalassemia variants has also become a common practice. The finding that specific β-thalassemia mutations were linked to patterns of restriction site polymorphism (haplotypes) in the human β-globin gene cluster was soon reported by Orkin *et al.* (1982). In this study, nine different patterns were described after assessing

cleavage at seven polymorphic sites in β-thalassemia chromosomes of patients of Mediterranean origin (Figure 5).

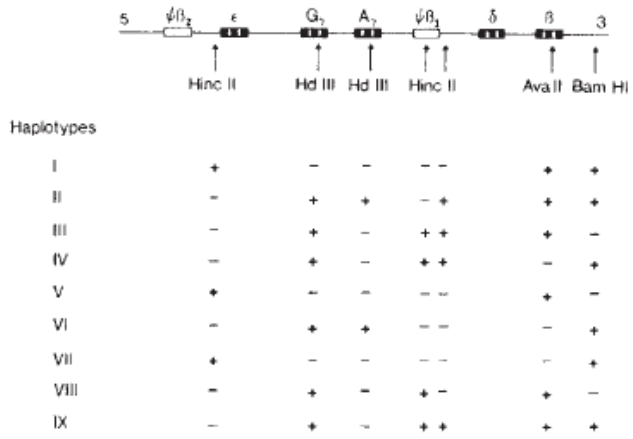


Figure 5. Schematic representation of the β-globin cluster containing the polymorphic restriction enzyme sites used and the cleavage patterns designated Haplotypes I-IX (adapted from Orkin *et al.* 1982).

Just like with SCD haplotypes, β-thalassemia haplotypes were also found to be related to different HbF levels. A study of 42 β-thalassemia major patients from Algeria has shown that on the one hand haplotypes IX (in haplotypic homozygotes and heterozygotes) and III (in heterozygotes) were linked to high ^{G}γ-globin expression, high levels of HbF and, consequently, patients with these haplotypes exhibited hematological amelioration of their disease and, on the other hand that haplotypes I and V were associated with low ^{G}γ-globin expression and low HbF levels (Labie *et al.* 1985). Labie *et al.* did also interrogate the *XmnI* site (5'Gγ), a site already screened for β^{S} haplotyping, because shortly before the T in the position -158 (detectable by *XmnI* digestion) had been reported to be associated with high ^{G}γ-globin expression both in sickle cell anemia and β-thalassemia patients (Guilman and Huisman, 1985). Importantly, the authors found out that the β-thalassemia haplotypes I and V were *XmnI*-negative. This finding led to conclude that most probably the *XmnI* position, later termed rs7482144, was indeed the responsible for promoting the expression of *HBG2*, contributing to HbF variability. Subsequent independent studies confirmed the association between *XmnI-HBG2* T allele and increased HbF, as well as milder disease among individuals with SCD and β-thalassemia from different populations (Thein *et al.* 2009).

RFLP based approaches continue to be the most used to perform haplotyping in the β-globin gene cluster. However, they rely in methods that are very labor-intensive and time-consuming (Sutton *et al.* 1989; Lanclos *et al.* 1991; Dimovski *et al.* 1991; Jackson *et al.* 1996; Vinson *et al.* 2004).

In an effort to expedite haplotyping, a few alternatives to the conventional methods are emerging, including a SNaPshot® Multiplex system that was implemented targeting the most informative positions defining the common β^S haplotypes, which was demonstrated to provide an accurate haplotyping of β^S chromosomes (Couto, 2017). The system contains six SNPs, all corresponding to the restriction sites conventionally examined in RFLP-based methods, four of which are shared to infer β^S and β-thalassemia haplotypes. Besides, that SNaPshot® system integrates two of the currently most accepted modifiers of HbF levels, which are rs7482144, located in the promoter region of *HBG2* and rs10128556, in the pseudogene *HBBP1.*

2. Objectives

Up to now, *BGLT3* and *HBBP1* have been barely studied and the relationship between genetic variations at these genes and levels of HbF in adults has been very scarcely addressed. Therefore, one of the objectives of this study, is to contribute to bring light to this matter, interrogating *BGLT3* and *HBBP1* in a sample of Portuguese β-thalassemia carriers, in whom the association between HbF levels and variations at *BCL11A*, *HMIP* and *HBG2* (promoter) has been previously assessed (Pereira *et al.* 2015), applying linear regression models to evaluate the relative effect of the identified loci as candidate to mediate the levels of HbF.

Since both *BGLT3* and *HBBP1* are located in one region of the β-cluster with strong LD, the study was also aimed at analyzing the influence of extended haplotypes in levels of HbF, taking advantage of a recently developed Multiplex SNaPshot® system for haplotyping the β-cluster.

3. Methodology

3.1. Sampling

In this study, 71 DNA samples from unrelated individuals of Portuguese origin (from the Coimbra region) with β-thalassemia minor were used. The same subjects already had been characterized for SNPs located in *BCL11A* (2p16), the intergenic region *HBS1L-MYB* (6q23), and the promoter of *HBG2* (11p15.5) by Pereira *et al.* 2015.

Informed consent was provided by all the participants. Data was available on HbF and HbA2 levels, previously determined by high performance liquid chromatography (HPLC) using Variant 2 (Pereira *et al.* 2015; Bio-Rad, CA, USA).

3.2. *BGLT3* Mutations Screening

3.2.1. PCR

To identify variations in the *BGLT3* gene, the entire region it encompasses (chromosome 11:5244554-5245546) was amplified by Polymerase Chain Reaction (PCR). Since the whole gene is approximately 1000bp long, to ensure maximum amplification yield and subsequent complete sequencing, two pairs of primers were selected: one of the forward primers was complementary to the 5'-flanking region of *BGLT3* and the other to an intermediate region, whereas one of the reverse primers was complementary to the 3'-flanking region of the gene and the other to an intermediate region.

The primers (Table 2) were designed according to the sequence of the *BGLT3* gene (ENST00000564523.2) available in the Ensembl platform (Yates *et al.* 2015). Afterwards, properties of the primers, such as melting temperature, G/C-content, and self-complementarity were analysed in the platforms Primer3 (Untergasser *et al.* 2012) and OligoCalc (Kibbe, 2007). To confirm the specificity of the binding of the primers to the target regions, the platform UCSC Genome browser – Blat (Kent, 2002), available at http://genome.ucsc.edu/, was used.

Primer	Sequence (5'>3')	Fragment Length (bp)
BGLT3 - F	GCC ACA AAC AAG AAA GAA TC	1137
BGLT3-R2	AAT AAA CAC CTC TAT CCA GC	
BGLT3 - IF	TCC TTG AAA TAC ACA TGG GG	619 (with BGLT3-R2)
BGLT3 - IR	CCC CAT GTG TAT TTC AAG GA	538 (with BGLT3-F)

Table 2. Primers for amplification of the *BGLT3* gene and expected fragment size.

The PCR reactions were prepared with 5µL of Quiagen® Master Mix Kit, 1 µL of each primer (2 µM), 2.5 µL of distilled water and 1.5 µL of DNA, making a total of 10 µL per reaction. Negative controls were performed by substituting the DNA with distilled water (RNase/DNase/Protease free). Amplification was then conducted in a T100™ Thermal Cycler (Bio-Rad) under the conditions described in Table 3.

Step	Temperature (°C)	Time	Cycles
Initial Denaturation	95	15 min	1
Denaturation	94	30 sec	
Annealing	52	1:30 min	35
Extension	72	2 min	
Final Extension	72	10 min	1
Hold	4	∞	

Table 3. PCR protocol for the amplification of the *BGLT3* gene.

Amplification products were then separated by horizontal polyacrylamide gel (acrylamide: bisacrylamide 19:1) electrophoresis using a Multiphor II Electrophoresis System (GE Healthcare) with MultiTemp III Thermostatic Circulator (Amersham Biosciences) and a Consort EV243 power supply. The running was performed at 180V with 0'GeneRuler 100 bp Plus DNA Ladder (Thermo Scientific™) and the results were visualized by the Silver Staining method according to the following sequential steps: 10 minutes in Ethanol (10%); 5 minutes in Nitric Acid (1%); two washings with distilled water; 20 minutes in Silver Nitrate (0.2%); two washings with distilled water; DNA visualization solution composed of 3g of Sodium Carbonate (0.28M), 1 mL of Formaldehyde (4%) and 100 mL of distilled water; 2 minutes in Acetic Acid for the termination of the reaction and a final washing with distilled water.

3.2.2 Automated DNA Sequencing

To prepare the amplified PCR products for sequencing, a purification step was performed which consisted of adding 0.5 µL of a mixture of ExoSAP-IT™ with FastAP with 1 µL of amplification product and then submitting the mixture to the conditions presented in Table 4. – Initial Purification.

Afterwards, to the previously purified sample (1.5 µL) a mixture of 0.8 µL BigDye™ Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems™), 1 µL BigDye™ Terminator v3.1 Sequencing Buffer (5x), 0.5 µL of primer (2.5 µM) and 1.2 µL of distilled water (RNase/DNase/Protease free) was added and then the samples were submitted to the conditions described in Table 4. – Sequencing Reaction.

Initial Purification

Temperature (°C)	Time
37	15 min
85	15 min

Sequencing Reaction

Temperature (°C)	Time	Cycles
96	2 min	1
96	30 sec	35
55	20 sec	
60	2 min	
30	10 min	1
Hold	∞	

Table 4. Protocol for sequencing of the amplified *BGLT3* fragments.

A final purification step by Sephadex™ G-50 Fine DNA Grade (GE Healthcare gel filtration resin was performed, following manufacture's protocol for the preparation of the columns. Then, after adding 8 µL of Formamide (Applied Biosystems™), the fragments were analysed in a 3500 Genetic Analyzer (Applied Biosystems™). Sequences were aligned and analysed using Geneious v5.5.6 analysis software (Biomatters, available at http://www.geneious.com).

3.3. SNaPshot® Multiplex System

For haplotyping the β cluster, it was followed the strategy recently implemented by Couto (2017), specifically directed to HBB*S haplotyping, that was based on a SNaPshot® Multiplex System. Herein, it will be described in detail the steps reported by Couto to develop and apply the system. First, an extensive bibliographic research on the haplotypes associated with sickle cell mutations was performed. Given that the conventional methodology for the identification of HBB*S haplotypes was based on Restriction Fragment Length Polymorphism (RFLP), the identification of sequences of the fragments that were typically amplified and then submitted to restriction enzymes digestion was done, followed by the determination of the exact location of the most informative polymorphisms defining the haplotypes, which resulted in the selection of six SNPs: rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905 (Figure 6).

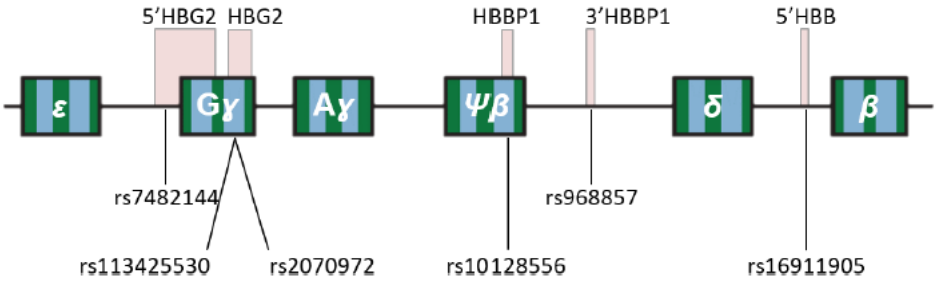


Figure 6. Schematic representation of the β-globin gene cluster, the target regions to be amplified (rose squares) enveloping the haplotype defining polymorphisms (Couto, 2017).

3.3.1 Multiplex PCR

To amplify the desired fragments, Couto designed five pairs of primers (Table 5) with different lengths, in order to implement a Multiplex PCR.

Primer	Sequence (5'>3')	Fragment Length (bp)
5'HBG2_FW	ACAAGAAGGTGAAAAACGG	772
5'HBG2_RV	CTTTATGGCATCTCCCAAG	
HBG2_FW	GCTGCAAGAAGAACAACTACC	400
HBG2_RV	GACAACCATGTGTGATCTCTTA	
HBBP1_FW	CAGGATTCTTTGTTATGAGTGTT	332
HBBP1_RV	CAAGCTGGACTTGCAGTAA	
3'HBBP1_FW	GAGACCTAAACTGAGGAACCTT	208
3'HBBP1_RV	CTTGATGGACCCTAACTGATATA	
5'HBB_FW	GATCACGTTGGGAAGCTATA	166
5'HBB_RV	AGGTCTTCTACTTGGCTCAGA	

Table 5. Primers for amplification of the target regions of the β-globin cluster and expected fragment sizes.

PCR reactions were performed in a final volume of 5 µL, containing 2.5 µL of Quiagen® Multiplex PCR Kit, 0.5 µL of Primer Mix (with a final concentration of 2 µM for each primer), 3.5 µL of distilled water (RNase/DNase/Protease free) and 0.5 µL of DNA, under the conditions in Table 6. Negative controls were performed by substituting the DNA with distilled water (RNase/DNase/Protease free). Amplification was conducted in a T100™ Thermal Cycler (Bio-Rad).

Step	Temperature (°C)	Time	Cycles
Initial Denaturation	95	15 min	1
Denaturation	94	30 sec	
Annealing	60	1:30 min	35
Extension	72	1 min	
Final Extension	72	10 min	1
Hold	4	∞	

Table 6. Multiplex PCR protocol for amplification of the target regions of the β-globin cluster.

The separation of the amplified products as well as the visualization of the DNA was as described in section 2.3.1.

3.3.2 SNaPshot®

The SNaPshot® Multiplex system relied on six Single Base Extension (SBE) primers, which were needed to simultaneously determine the genotypes at the six target SNPs. The SBEs were designed to hybridize with the sequence immediately adjacent to the

target nucleotide base. However, to turn possible the discrimination (by size) of the different primers after submitted to capillary electrophoresis, non-annealing tails with different sizes were added to each of the six SBE primers, presented in Table 7.

rs	Polymorphism	Detected Allele	Sequence (5'>3')
7482144	G>A	G/A	GGTGGAGTTTAGCCAGG
113425530	C>A	G/T	GTCGTGAAAGTCTGACAATTGATTCTGGGTG GAA
2070972	A>C	A/C	GACTAAACTAGGTGCCACGTCGTGAAAGTCT GACAACCTCCAGATAACTACACACC
10128556	C>T	G/A	GTCTGACAATGTTGGGGTAGTGAGTTG
968857	T>C	T/C	CAATGCATGACACATGCTTG
16911905	G>C	G/C	TGCCACGTCGTGAAAGTCTGACAACGTTTTA AAATCATTTCCTT

Table 7. Single Base Extension primers for the genotyping of the target polymorphisms.

In order to optimize the quality of the samples for the SNaPshot® reactions, before them a purification step was performed, that consisted in adding 0.5 µL of ExoSAP-IT™ (Applied Biosystems™) to 1 µL of amplification product and then submitting the solutions to the conditions presented in Table 8 – Initial Purification.

Then, to each purified sample, 1 µL of SNaPshot™ Multiplex Kit (Applied Biosystems™), 1 µL of SBE Primer Mix and 1.5 µL of distilled water (RNase/DNase/Protease free) was added. The SBE concentrations needed to obtain the SBE Primer Mix are detailed in Table 9. The SNaPshot® reactions took place under the conditions described in Table 8 - SNaPshot®.

Initial Purification		SNaPshot®			Final Purification	
Temperature (◦C)	**Time**	**Temperature (◦C)**	**Time**	**Cycles**	**Temperature (◦C)**	**Time**
37	15 min	96	10 sec		37	1 h
85	15 min	50	5 sec	25	85	15 min
		60	30 sec			

Table 8. Protocols for the purification steps and genotyping of the target polymorphisms.

A final purification step (Table 8 – Final Purification) was performed by adding 1 µL of FastAP (Thermo Scientific™) to the final product.

The described reactions were all conducted in a T100™ Thermal Cycler (Bio-Rad).

SBE	Concentration in final solution (5 µL)
rs7482144	0.8 µM
rs113425530	1.6 µM
rs2070972	0.5 µM
rs10128556	0.3 µM
rs968857	0.6 µM
rs16911905	0.6 µM

Table 9. SBE concentrations used to prepare the SBE Primer Mix.

In order to analyze the products by capillary electrophoresis, in a 3500 Genetic Analyzer (Applied Biosystems™), 10 µL of a mixture of Hi-Di™ Formamide (Applied Biosystems™) and GeneScan™ – 120 LIZ™ Size Standard (Applied Biosystems™) were added to 0.53 µL of the SNaPshot® purified products. The results were analyzed with the GeneMapper® Software 5.

3.3.3 Conventional HBB*S Haplotypes

The non-allele combinations that define the five main haplotypes, Bantu/CAR, Benin, Senegal, Arab-Indian and Cameroon, are presented in Table 10.

	rs7482144	rs113425530	rs2070972	rs10128556	rs968857	rs16911905
Bantu/CAR	G	G	A	G	C	C
Benin	G	T	C	G	T	C
	G	T	A	G	T	C
	G	G	C	G	T	C
Senegal	A	G	A	A	T	G
Arab-Indian	A	G	A	A	T	C
Camaroon	G	G	A	G	T	G

Table 10. Single Nucleotide Polymorphisms alleles associated to the main sickle cell haplotypes.

As explained by Couto, once rs113425530 and rs2070972 are located in the same restriction enzyme recognition sequence, to establish the correspondence with results obtained with RFLP analysis (cut/+; absence of cut/-) the genotypes at the two variant sites must be analyzed as a single result whereby the presence of nucleotide T in the first SNP or C in the second was assumed to promote enzyme recognition, while non-recognition was presumed to occur in the presence of G and A in rs113425530 and

rs2070972, respectively. For that reason, three different SNP-inferred haplotypes fell in the RFLP-based Benin configuration.

3.4 Statistical Analysis

Allele frequencies of the polymorphisms were estimated by direct counting. Associations of SNPs with HbF levels were performed, after logarithmic transformation in order to near normalize the quantitative trait distribution, estimating P values and 95% confidence intervals (CI), crude and with age and sex as covariates. All of these analyses were performed using the tests implemented on PLINK software v.1.07 (available at http://zzz.bwh.harvard.edu/plink/) (Purcell *et al.* 2007). This software was also used to determine LD values between the SNPs studied. In order to obtain LD blocks, Haploview (Barrett *et al.* 2005) v.4.2 was used.

Moreover, graphical analyses, normality of the data assessed by the Kolmogorov-Smirnov test and comparisons of HbF levels between genotypes using the Kruskal-Wallis (1-way-ANOVA) test were performed with IBM® SPSS® Statistics v.26 software.

Haplotype inference was performed with the software Arlequin v.3.5 (Excoffier and Lischer, 2010).

Network Analysis was executed with the Network v.5.0 software (available at http://www.fluxus-engineering.com/).

4. Results and Discussion

A compilation of demographic characteristics and hematological parameters of the individuals analyzed in this study is presented in Table 11.

Characteristics	Sampled individuals
Age (mean ± SD, in years)	32,56 (± 20,5)
Age (range, in years)	2 -- 77
Males (n)	37
Females (n)	34
HbF (mean ± SD; %)	1,78 (± 1,76)
HbF (range; %)	0,2 -- 8,6
HbA2 (mean ± SD; %)	4,76 (± 0,67)
HbA2 (range; %)	3,4 -- 6,1

Table 11. Demographic and hematological data from the studied subjects.

All the 71 subjects had β-thalassemia minor. They were heterozygous for seven distinct β-thalassemia mutations, which were: : IVSI-6 T>C (c.92+6T>C), carried by 27 subjects; CD39 (CAG>TAG) (c.118C>T), present in 22 subjects; CD 15 TGG>TGA (c.48G>A), 8 subjects; IVSI-1 G>A (c.92 + 1G>A), 7 subjects; CD 15 (TGG>TAG) (c.47G>A), 3 subjects; IVSI-110 G>A (c.93-21G>A), 2 subjects; and CD 6 -A (c.20delA), 2 subjects.

4.1 *BGLT3* Molecular Screening

In view of the genetic characterization of *BGLT3*, a direct sequencing approach was applied encompassing the entire sequence length of the gene. To obtain accurate sequencing results, it was necessary to ensure clear amplifications, which were successfully achieved after optimizing the correct amount of DNA needed and the appropriate PCR conditions. The amplification products revealed the absence of non-specific bands, which critically could interfere with the results, while the yield of the desired amplicons was very good. In Figure 7 is depicted an example of the results obtained after submission of the PCR products to electrophoretic separation.

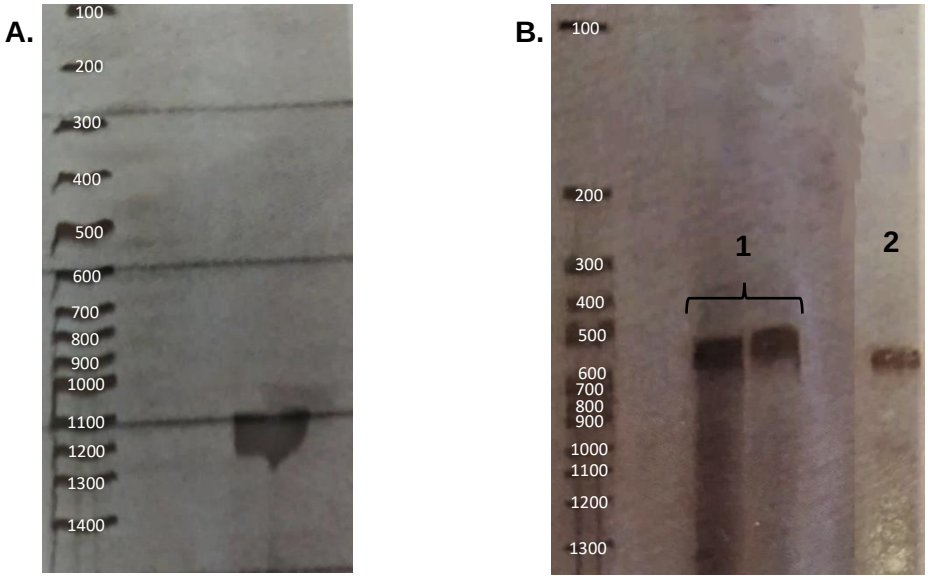


Figure 7. Gels showing PCR products amplified with **A.** primers BGLT3-F and BGLT3-R2; **B.1** primers BGLT3-F and BGLT3-IR and **B.2** primers BGLT3-IF and BGLT3-R2.

The sequencing reactions also led to produce good and consistent sequencing data as is shown in Figure 8. The results indicated that position 49 of the *BGLT3* gene (chromosome 11:5245498) was highly polymorphic in the sample analyzed. Of the 71 subjects examined, 56 had the T allele at that position, out of which 19 in homozygosity and the remaining 37 in heterozygosity (see in Figure 8 chromatograms illustrating the different genotypes observed for this polymorphism and Table S1 for more detailed information).

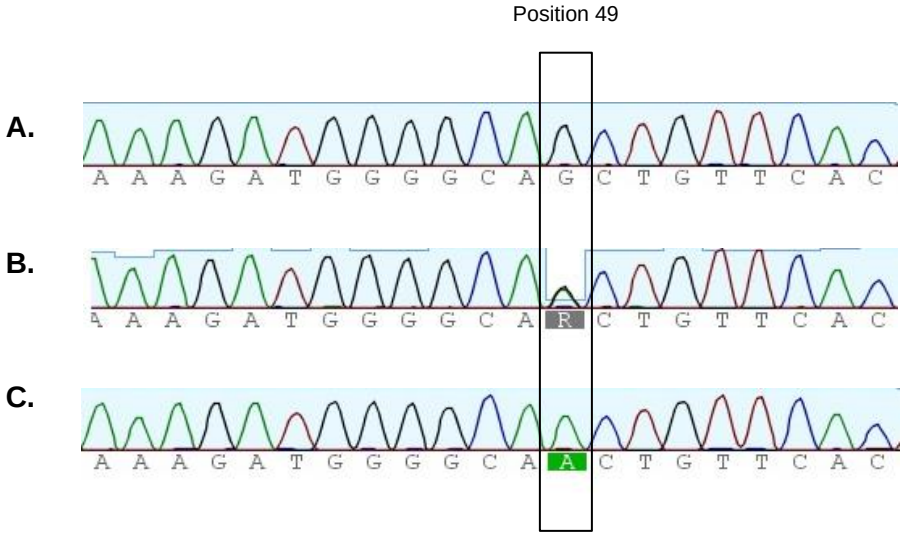


Figure 8. Examples of sequencing results for the rs7924684 SNP in the BGLT3 gene. **A.** Homozygous G/G; **B.** Heterozygous G/A; **C.** Homozygous A/A.

In the Ensembl platform, this SNP (rs7924684) is described as a C>T non-coding transcript exon variant, with the derived allele (T) appearing widespread in all human populations, though being especially common in non-African ones; for instance, in the European super-population it attains the frequency of 49%. In the sample here studied the frequency of the T allele was 0.472, which fits very well in the range of values usually found in European populations. Regarding the genotypic distribution for rs7924684 observed in the sample, no departures from HW expectations were detected (Table 12; p_{HW} =0,814).

Besides rs7924684, three other variations were found in the *BGLT3* gene, namely rs76914448, rs568267053 and rs917484303. According to the Ensembl platform, all of them are non-coding transcript exon variants. The first one is an A>G mutation located in chromosome 11:5245003, the second a G>A in chromosome 11:5244763, and the third a G>A substitution located in chromosome 11:5244763. All the three derived alleles are extremely rare in European populations. In this study each of them was detected in a single different heterozygous individual. These 3 subjects presented intermediate values of HbF (2,9 – 4,6).

Due to very low frequency of the mutated alleles at the 3 SNPs in the sample, they were not be submitted to association analysis with levels of HbF.

SNP	N	Alleles (1:2)	MAF	P-HWE
rs16911905	71	C:G	0,092	1
rs968857	71	T:C	0,43	0,636
rs10128556	71	A:G	0,204	0,465
rs7924684	71	G:A	0,472	0,814
rs2070972	71	C:A	0,345	0,007
rs113425530	71	T:G	0,007	1
rs7482144	71	T:C	0,197	0,447

Table 12. Minor allele frequencies (MAF) and Hardy-Weinberg Equilibrium assessment (P-HWE) for each SNP.

4.2. Data provided by the SNaPshot® Multiplex System

As previously mentioned, in order to haplotype the β cluster, a strategy based on a SNaPshot® Multiplex System, developed by Couto (2017), was used. This multiplex

system targets the six SNPs demonstrated to be suitable for ascertain the conventional RFLP-based HB*S haplotypes, which were: rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905. A representation of an electrophoretic run after the Multiplex PCR as well as electropherograms plotting the SNaPshot® results from two subjects are presented in Figure 9 and Figure 10, respectively. Complete genotypic data is presented in Table S1.

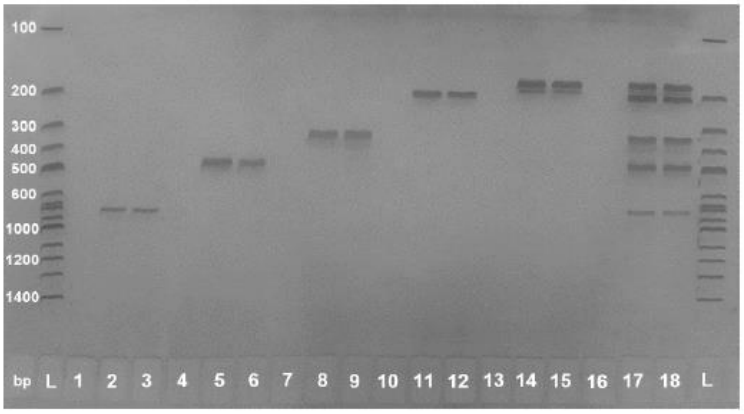


Figure 9. Gel showing the results of the Multiplex PCR targeting 6 SNPs. **L** – Ladder; **1** – 5'HBG2 Control; **2,3** – 5'HBG2; **4** – HBG2 Control; **5,6** – HBG2; **7** – HBBP1 Control; **8,9** – HBBP1; **10** – 3'HBBP1 Control; **11,12** – 3'HBBP1; **13** – 5'HBB Control; **14,15** – 5'HBB; **16** – Multiplex PCR Control; **17,18** – Multiplex PCR.

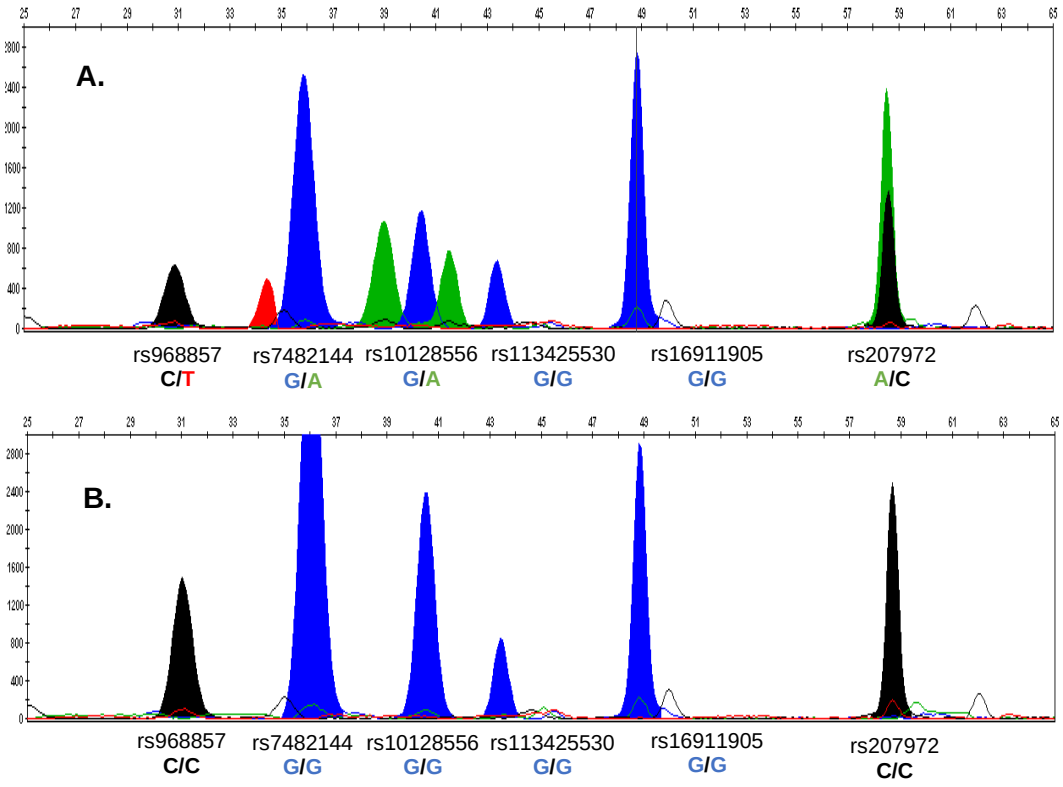


Figure 10. Examples of SNaPshot® reaction results and prediction of genotypes associated with each variant. **A.** sample 538; **B.** sample 458.

For each of the 6 SNPs interrogated with the SNaPshot® Multiplex, minor allele frequencies (MAF) and the P-values for the Hardy-Weinberg equilibrium (HWE) tests are also presented in Table 12. The allele frequencies in our sample fell in the typical range observed in European populations and all genotypic distributions were in agreement with predicted under HWE when a *P*-value adjusted, with the Bonferroni correction (Armstrong, 2014), for multiple tests was assumed.

From the genotypic data for the 6 SNPs, haplotypes were estimated with the Expectation Maximization/Excoffier-Laval-Balding (EM/ELB) algorithm implemented in Arlequin 3.5.2.2v software. EM/ELB algorithms make an initial guess of the haplotype frequencies and they are able to assign all alleles to haplotypes with a high probability; assuming the data are in Hardy-Weinberg equilibrium (Crawford *et al.* 2005).

A total of ten haplotypes were inferred in the sample (Table 13).

Haplotype ID	Haplotype	Frequency	Observations
1	AGAATG	26	Senegal
2	GGAGCG	29	
3	GGAGTG	22	Cameroon
4	GGCGCG	49	
5	GGAGTC	10	
6	GGAGCC	2	Bantu/CAR
7	AGAACG	1	
8	GGAATG	1	
9	GTAGTG	1	
10	AGAATC	1	Arab-Indian

Table 13. Inferred haplotypes associated to the 71 studied samples, by Arlequin software along with frequency and identification of the main haplotype associated. Haplotypes defined by polymorphisms in the following order: rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905.

Four of them corresponded to some of the conventional HBB*S haplotypes (section 3.4 – Table 10), namely haplotypes 1, 3, 6 and 10, which matched the haplotypes Senegal, Cameroon, Bantu/CAR and Arab-Indian, respectively. In total, they summed up 51 occurrences, representing 35.9% of the entire set of haplotypes.

From the estimates of the gametic phase for each individual, no evidence emerged that the presence of any of the four conventional haplotypes was related with a particular β-thalassemia mutation. For instance, the SEN or CAM haplotypes, the two most well represented in the sample, were indiscriminately carried by subjects with different β-

thalassemia mutations. Obviously that no other result was expected, since these haplotypes are known to be alternative backgrounds only of HBB*S alleles.

However, it was possible to clearly discriminate the distinct non-conventional haplotypes (or groups of related non-conventional haplotypes) to which were anchored to the 7 different β-thalassemia mutations, as will be presented in the next section.

4.3 Linkage Disequilibrium Analysis

Considering the data for the SNPs included on the SNaPshot® Multiplex System together with that for the highly polymorphic position detected in the *BGLT3* gene, we end up with genotypic data for seven SNPs located throughout a region encompassing a great extension of the entire β-globin cluster. This cluster contains two distinct regions with strong LD: one that contains *HBB* and the other extending from *HBD* to the LCR, with the two regions being separated by a segment that constitutes a recombination hotspot in the cluster (see, for instance, Moleirinho *et al.* 2013). In Figure 11 is presented a map of the LD pattern in the β-globin cluster, where are indicated the approximate positions of seven SNPs here analyzed.

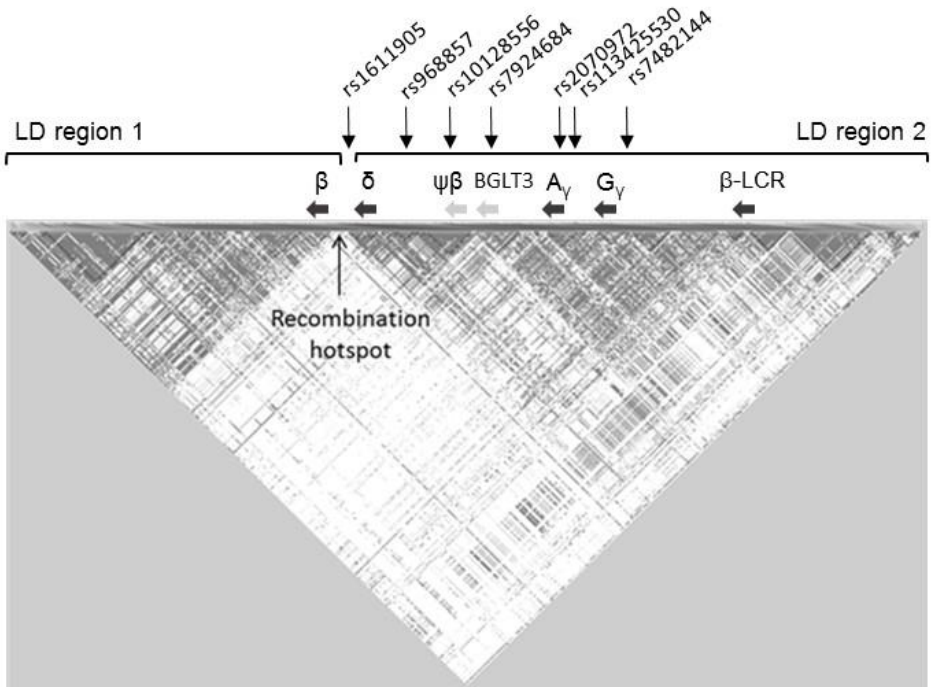


Figure 11. LD plot of the the β-globin cluster for the data from 1000 Genomes Project for CEU (adapted from Molerinho *et al*. 2013) and location of the set of SPNs analysed in this study.

The LD analysis here performed with the seven SNPs revealed a pattern of LD that is illustrated in Figure 12. A well-defined LD block was observed that involved 4 SNPs: rs968857, rs10128556, rs7924684 located in the intergenic region between *HBBP1* and *HBD*, and rs2070972, located in *HBG2.* Within this block, two SNPs revealed to be in almost total LD that were rs968857 and rs7924684, the first located in the intergenic region between *HBBP1* and *HBD*, and the second in *BGLT3*. Remarkably, one of the SNPs from this block, namely rs10128556, located in the pseudogene *HBBP1*, was also in very strong LD with a SNP not belonging to the block, which was rs7482144, located in the promoter region of *HBG2*. Taking into account the relative positions of the SNPs in the β-globin cluster, the observed pattern of LD is rather weird suggesting that besides bp distance between each pair of SNPs, interactions between SNPs must also account for the level of non-random association of alleles at different SPNs.

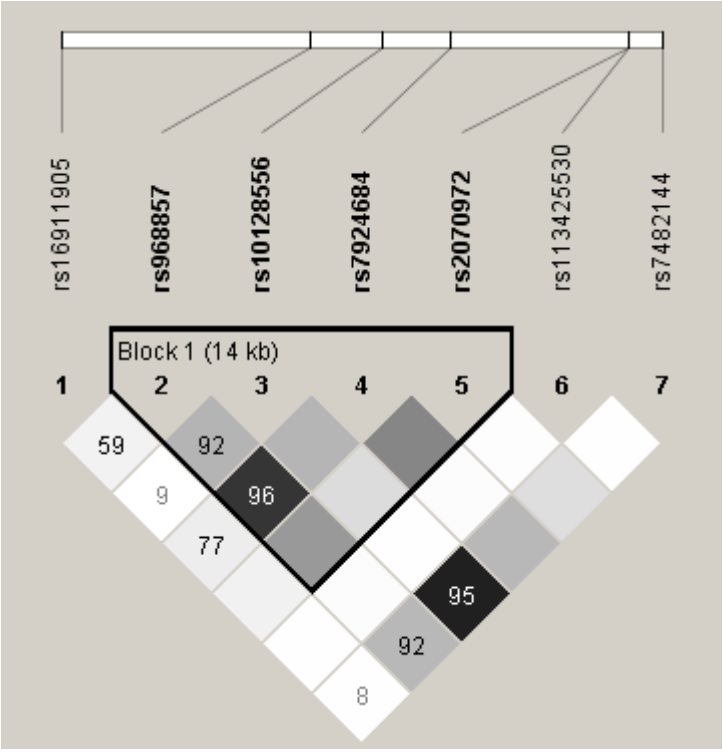


Figure 12. Schematic representation of LD Blocks created for the pairs of SNPs, according to their LDs. The values appearing inside the squares represent the D' values. Squares with no number on the inside have D' values of 1.

Concerning the haplotypes defined by the seven loci, thirteen different haplotypes were inferred in our sample, some of which being in strong association with β-thalassemia mutations. A network analysis was then performed to facilitate the description of the results obtained (Figure 13).

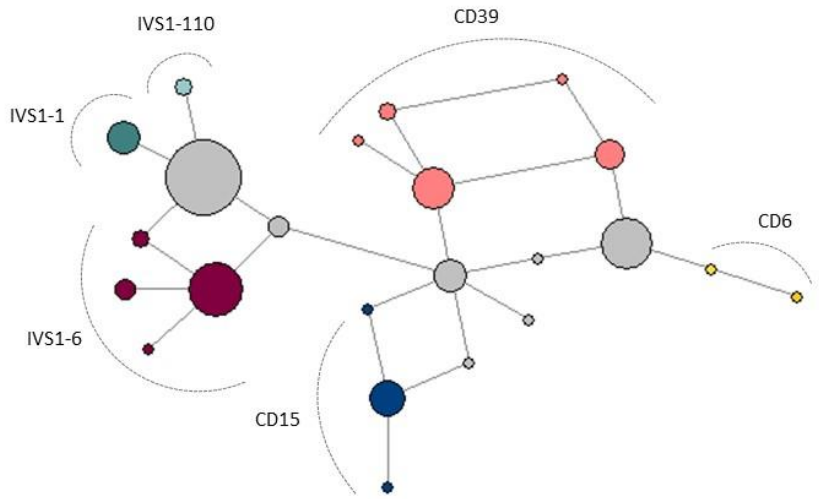


Figure 13. Network of haplotypes defined by 7 SNPs spread in the β-globin cluster plus β-thalassemia mutations. The gray circles represent the haplotypes that do not have any β-thalassemia mutation.

The most common β-thalassemia mutation in our sample was IVSI-6, whose core seven-SNPs haplotype was GGAAGCG, considering the following SNP order: rs7482144, rs113425530, rs2070972, rs7924684, rs10128556, rs968857 and rs1691195. From this core haplotype, three other single step derived haplotypes are also anchored to IVSI-6. The level of (SNP inferred) haplotypic diversity within IVS1-6 chromosomes seems high enough to suggest that this β-thalassemia mutation must be rather old. The same holds for the second most common β-thalassemia mutation in the sample, CD39. It was possible to identify a core haplotype, GGAGGTG, which shows close phylogenetic relationships with the remaining 4 haplotypic backgrounds found in CD39 chromosomes.

The third commonest mutation, CD15, was present in 3 related haplotypes, of which GGAGGTC was the core.

Even taking into account both the complex evolutionary dynamic of the entire β-globin cluster and the presence of a recombination hot spot separating the HBB from the SNPs

here screened, the level of haplotypic diversity within IVSI-6, CD39 and CD15 seems difficult to explain assuming a relatively recent origin.

4.4 Association Analysis with HbF Levels

To address the association between the screened SNPs and levels of HbF, linear regression analysis was performed, first individually for each locus and next considering the multi-loci inferred haplotypes. In any case, the analyses were carried under an additive model and assessing the distribution of log transformed HbF levels.

In Table 14 are presented the association results obtained individually for each SNP.

Chr: position	SNP	N	β(SE)	P	P*
11:5228060	rs16911905	71	0,374 (0,307)	0,227	0,085
11:5239228	rs968857	71	0,7 (0,144)	**$7{,}091 \times 10^{-6}$**	**$5{,}342 \times 10^{-6}$**
11:5242453	rs10128556	71	0,65 (0,185)	**$8{,}092 \times 10^{-4}$**	**$1{,}845 \times 10^{-3}$**
11:5245498	rs7924684	71	0,716 (0,151)	**$1{,}095 \times 10^{-5}$**	**$8{,}250 \times 10^{-6}$**
11:5253487	rs2070972	71	-0,258 (0,217)	0,239	0,1342
11:5253490	rs113425530	71	-0,752 (1,013)	0,46	0,4574
11:5254939	rs7482144	71	0,685 (0,185)	**$4{,}096 \times 10^{-4}$**	**$6{,}499 \times 10^{-4}$**

Table 14. HbF association results in individuals with β-thalassemia minor of Portuguese origin.

The table includes the effect sizes of the minor allele (regression coefficient beta, β), standard error (SE), P-values for the log-transformed HbF levels (P) and P-values using age and sex as covariates (P*).
Chromosome position is according to Ensembl.
Significant association P-values ($P < 0.05$) are in bold.

Four of the seven SNPs revealed to be significantly associated with HbF levels: rs7482144, rs7924684, rs10128556 and rs968857, located in the promoter region of *HBG2*, in *BGLT3*, in *HBBP1* and in the intergenic region between *HBBP1* and *HBD*, respectively. Using age and sex as covariates did not significantly altered the results. For all these four SNPs, the yielded regression coefficients β were astonishingly high (>0.64) and accordingly the associated P-values were extremely low ($P< 4.1 \times 10^{-4}$).

The strongest signal was shown by rs968857 ($P= 7.091 \times 10^{-6}$), which is positioned in a DNA fragment that involves the Corfu deletion. This is a 7.2kb deletion between *HBBP1*

and *HBD* that typically is connected with elevated HbF levels, supposedly because it contains binding sites for BCL11A (Galanello et al, 1990, Kulozik et al 1988).

Recently, a fine sequence examination of the region encompassing the Corfu deletion in patients with SCD, also led to identify rs968857 as being associated with elevated HbF levels (Akinsheye *et al.* 2011). Furthermore, in silico analysis revealed that the T-C polymorphism at rs968857 eliminates NF-E2 and AP-1 (another transcription factor involved in the hemoglobin switch mechanisms), which are present in the minor allele, although the functional significance of this SNP remains to be explored (Akinsheye *et al.* 2011). Interestingly, among the conventional HBB*S haplotypes the unique one that harbors the major allele at this SNP is the Bantu/CAR, which is associated with low levels of HbF and the more severe SCD phenotype (Leal *et al.* 2016; Steinberg, 2009).

Regarding rs7482144, located at the promoter region of the *HBG2* gene, it also showed to be highly associated with HbF levels. The result was quite predictable since rs7482144, often referred to as *HBG2* (XmnI), represents one of the first recognized fetal hemoglobin QTL (Galarneau *et al.* 2010; Nuinoon *et al.* 2010). Furthermore, Pereira *et al.* 2015, did perform a previous study based essentially in the same sample of β-thalassemia carriers here used, where rs7482144 had already been investigated together with six more SNPs located at BCL11A intron-2, HBS1L-MYB (HMIP) intergenic region and KLF1. If in the study of Pereira *et al.*, out of the seven SNPs examined, rs7482144 revealed by far the strongest association with HbF (β=0.455; $P=5.858 \times 10^{-7}$), in the current work based in the very same sample plus a few more carriers, the result was clearly replicated (β=0.685; $P=4.096 \times 10^{-4}$), but it only represented the third greatest association detected.

The fourth strongest implied rs10128556 ($P = 8.092 \times 10^{-4}$), a SNP residing within the second intron of the *HBBP1* pseudogene. Although being one of the SNPs that defines the conventional HBB*S haplotypes, convincing evidence that it influenced HbF levels only became to emerge very recently through genome wide association studies (Galarneau *et al.* 2010; Nuinoon *et al.* 2010). Our results on rs10128556 clearly add strength to these previous findings. The detection of SNPs in *HBBP1* (besides rs10128556 also rs2071348, which actually are in total LD in most populations) associated with elevated levels of HbF sustains the suggestion that *HBBP1* contains developmental regulatory elements (Huang et al 2017).

A major problem when dealing with SNPs from the β-globin cluster is its intricate pattern of LD, and specifically regarding rs10128556 (in *HBBP1*) and rs7482144 (in the promoter region of the *HBG2*) our LD analysis revealed that they are almost in total LD, as

presented in the last section. Thus, it is extremely difficult to assess the real role of each of the two SNPs in the modulation of HbF levels (see section 4.3).

Consistently, however, the two conventional HBB*S haplotypes that contain the A alleles at both these SNPs, the SEN and AI haplotypes, are associated with high HbF levels and thus with the less severe SCD phenotypes.

Much more unexpected was to find out that, in our samples, the second most significant association results came from rs7924684, located in the *BGLT3* gene ($P = 1.095 \times 10^{-5}$). Although this gene has been very recently implicated in γ-globin expression (Ivaldi *et al.* 2018), until now there was no evidence of any specific SNP in *BGLT3* related to increased HbF levels. Further studies must be done, in order to reinforce these results, preferentially with samples enriched in particular β-thalassemia mutations to avoid any bias that mutation heterogeneity in *HBB* might induce. In addition, association studies with HbF levels from healthy individuals are as well urgently needed.

The distribution of log-transformed HbF values according to SNP genotypes was analyzed by the Kruskal-Wallis (1-way-ANOVA) test using IBM® SPSS® Statistics v.26 software. This non-parametric test was the one implemented here because, after the Kolmogorov-Smirnov normality test it was established that the data was not normally distributed. Results are summarized in Table 15.

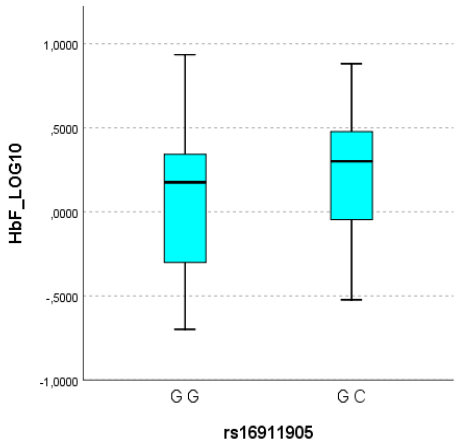
SNP	Genotypes HbF Mean (SD)			P*
rs16911905	**GG** (n = 58) 1,83 (0,22)	**GC** (n = 13) 2,48 (0,61)	**CC** (n = 0) -	0,264
rs7924684	**GG** (n = 15) 3,09 (0,53)	**GA** (n = 37) 2,05 (0,29)	**AA** (n = 19) 0,86 (0,20)	**<0,001**
rs968857	**CC** (n = 24) 0,93 (0,17)	**CT** (n = 33) 2,17 (0,32)	**TT** (n = 14) 3,2 (0,56)	**<0,001**
rs10128556	**GG** (n = 46) 1,56 (0,25)	**GA** (n = 21) 2,49 (0,38)	**AA** (n = 4) 3,63 (0,97)	**0,004**
rs2070972	**AA** (n = 25) 2,34 (0,41)	**AC** (n = 43) 1,80 (0,26)	**CC** (n = 3) 0,87 (0,47)	0,375
rs113425530	**GG** (n = 70) 1,97 (0,22)	**GT** (n = 1) -	**TT** (n = 0) -	0,434
rs7482144	**CC** (n = 47) 1,53 (0,24)	**CT** (n = 20) 2,61 (0,39)	**TT** (n = 4) 3,63 (0,97)	**0,002**

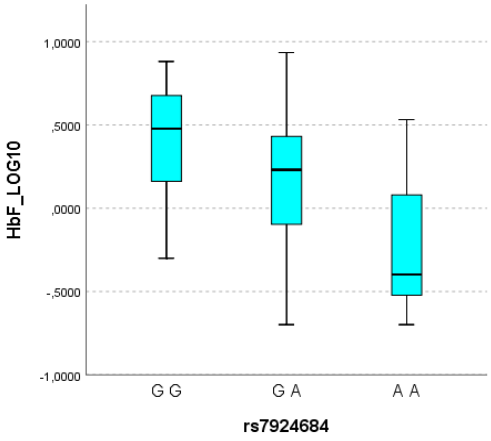
Table 15. Associations of the HbF levels (log-transformed) with genotypes of the studied SNPs in 71 individuals with β-thalassemia minor.

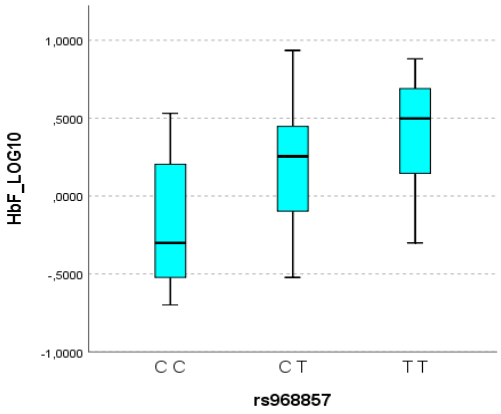
n: number of subjects within each genotype

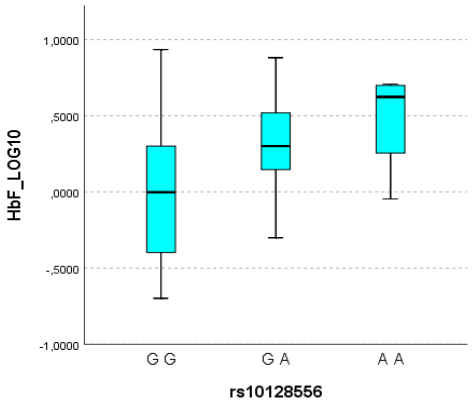
P*: P-value obtained with the Kruskal-Wallis (1-way-ANOVA) test. Significant P-values (P<0.05) are in bold.

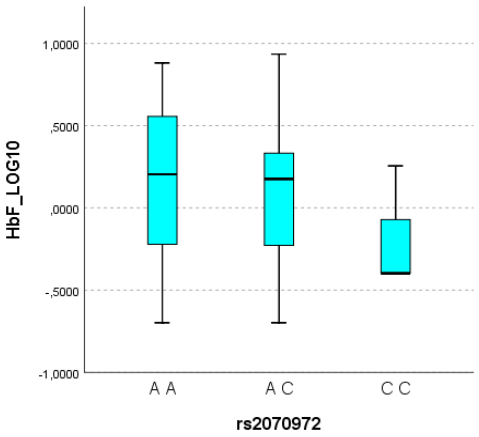
Significant statistical differences were observed for rs7482144 (P = 0.002), rs7924684 (P < 0.001), rs10128556 (P = 0.004) and rs968857 (P < 0.001), which is in accordance with the association test previously presented. In all cases, individuals homozygous for the minor allele have HbF levels about three times higher than those homozygous for the major allele. Figure 14 shows the box-plots displaying the variation on our data set.

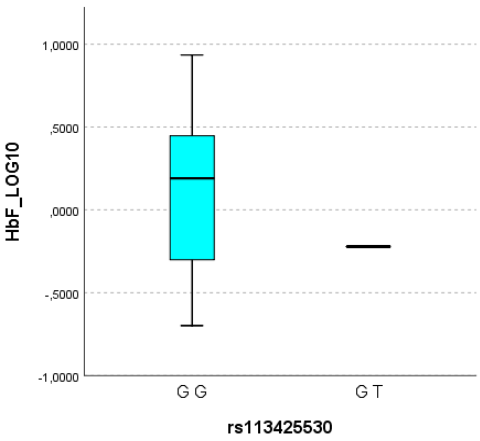

HbF_LOG10
1,0000
,5000
,0000
-,5000
-1,0000
G G
G C
rs16911905


HbF_LOG10
1,0000
,5000
,0000
-,5000
-1,0000
G G
G A
A A
rs7924684


HbF_LOG10
1,0000
,5000
,0000
-,5000
-1,0000
C C
C T
T T
rs968857


HbF_LOG10
1,0000
,5000
,0000
-,5000
-1,0000
G G
G A
A A
rs10128556


HbF_LOG10
1,0000
,5000
,0000
-,5000
-1,0000
A A
A C
C C
rs2070972


HbF_LOG10
1,0000
,5000
,0000
-,5000
-1,0000
G G
G T
rs113425530

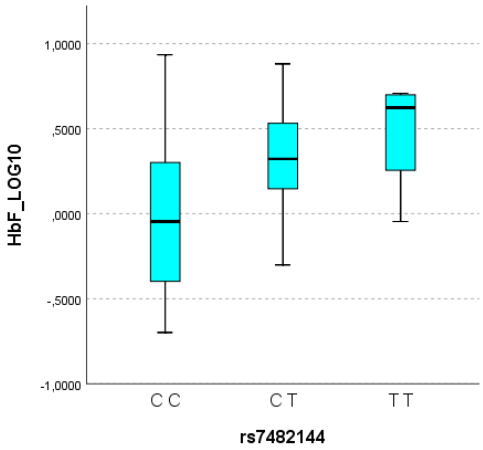


Figure 14. Box-plots showing the distribution of log-transformed HbF levels within genotypes of the SNPs rs16911905, 7924684, rs968857, rs10128556, rs2070972, rs113425530 and rs7482144 in Portuguese individuals with β-thalassemia minor. Each rectangle represents the data between the 25th and the 75th quartiles and the bar across each rectangle is the median value for HbF.

Given the pattern of LD between the seven SNPs under study, besides assessing the association between HbF levels at each locus, the association between HbF levels and multi-loci inferred haplotypes was also addressed.

At first, the results based only on the SNPs targeted by the SNaPshot® Multiplex System (rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905) will be presented, given the widely recognition that each of the conventional HBB*S haplotypes is associated with a characteristic average level of HbF, specifically that the SEN and AI haplotypes are associated with the highest levels of HbF; the BEN and CAM are related to intermediate HbF levels and the Bantu/CAR haplotype is correlated with the lowest levels of HbF (Powars, 1991; Steinberg, 2009; Crawford *et al.* 2002; Leal *et al.* 2016; Hirokawa *et al.* 1995) in patients with SCD, and so, it appeared interesting to assess how those six SNP-defined haplotypes were related with HbF levels in patients with β-thalassemia minor. In Table 16 are shown the results concerning haplotypes detected more than once in the sample.

Haplotype	Observations	β	P
AGAATG	Senegal	0,7862	**1,429 x 10^{-4}**
GGCGCG		-0,2577	0,2394
GGAGTC		0,6408	0,09215
GGAGTG	Cameroon	0,2616	0,287
GGAGCC	Bantu/CAR	-0,4805	0,5057
GGAGCG		-0,822	**3,675 x 10^{-5}**

Table 16. Associations of the HbF levels (log-transformed) with haplotypes inferred with the SNaPshot® Multiplex System.
β (regression coefficient beta): effect sizes of the minor allele

All shown haplotypes had a frequency greater than 1%.
Haplotypes were as follows: rs7482144 | rs113425530 | rs2070972 | rs10128556 | rs968857 | rs16911905.
Significant association P-values (P < 0.05) are in bold.

Among the most common six SNP-defined haplotypes, two revealed significant association with high HbF levels. One showed a positive correlation, AGAATG (P = 1.378 x 10^{-4}), which corresponds to the SEN haplotype, result that is in agreement with previous reports on the association of this haplotype with increased HbF levels (Leal *et al.* 2016; Steinberg, 2009). The other presented the highest negative correlation and was GGAGCG (P = 3.675 x 10^{-5}), which didn't correspond to any of the conventional HBB*S haplotypes. However, it only differed from the Bantu/CAR haplotype (also associated with low HbF levels) in rs16911905, which was a SNP that individually did not reveal to affect HbF. Therefore, it was likely to admit that within this haplotype could be hidden further variation relevant to modulate HbF.

To address the issue, identical association analysis was performed with the six SNPs targeted by the SNaPshot® Multiplex System plus rs7924684, in the *BGLT3* gene. The results are shown in Table 17, and again only for haplotypes with frequency >1%.

Haplotype	Observations	β	P
AGAGATG	Senegal + rs7924684 G	0,7864	**1,378 x 10^{-4}**
GGAGGTG	Cameroon + rs7924684 G	0,2441	0,3254
GGAAGCG		-0,8055	**8,444 x 10^{-5}**
GGCAGCG		-0,2577	0,2394
GGAGGCG		-0,2955	0,5594
GGAGGTC		0,6654	0,0879

Table 17. Associations of the HbF levels (log-transformed) with haplotypes inferred with the SNaPshot® Multiplex System plus the *BGLT3* gene polymorphism rs7924684. **β** (regression coefficient beta): effect sizes of the minor allele

All shown haplotypes had a frequency greater than 1%.
Haplotypes were as follows: rs7482144 | rs113425530 | rs2070972 | rs7924684 | rs10128556 | rs968857 | rs16911905.
Significant association P-values (P < 0.05) are in bold.

Two highly significant associations were detected, both encompassing the six SNP-defined haplotypes equally showing the strongest signals in the previous analysis. Among the seven-SNP haplotypes, the most significant association was GGAAGCG, the haplotypic background that doesn't correspond to any of the conventional HBB*S haplotypes plus the major allele (A) for the SNP in *BGLT3* (P = 8.444 x 10^{-5}), which was strongly associated with decreased HbF levels. Note that haplotype GGAGGCG, which differs from the former just by possessing G instead of A in the *BGLT3* SNP, was not associated with HbF levels.

The other statistically significant result involved the SEN haplotype combined with the G allele at rs7924684, AGAGATG (P = 1.378 x 10^{-4}), which accounted significantly to increased HbF levels.

These two seven-SNP haplotypes that highly associated with HbF levels, are quite different (Figure 15) and remarkably present alternative alleles at the four positions that in the SNP by SNP analysis showed highly significant associations: rs7482144, rs7924684, rs101288556 and rs968857. Among these four SNPs, while the first was found in strong LD with the third, the second was in strong LD with the fourth.

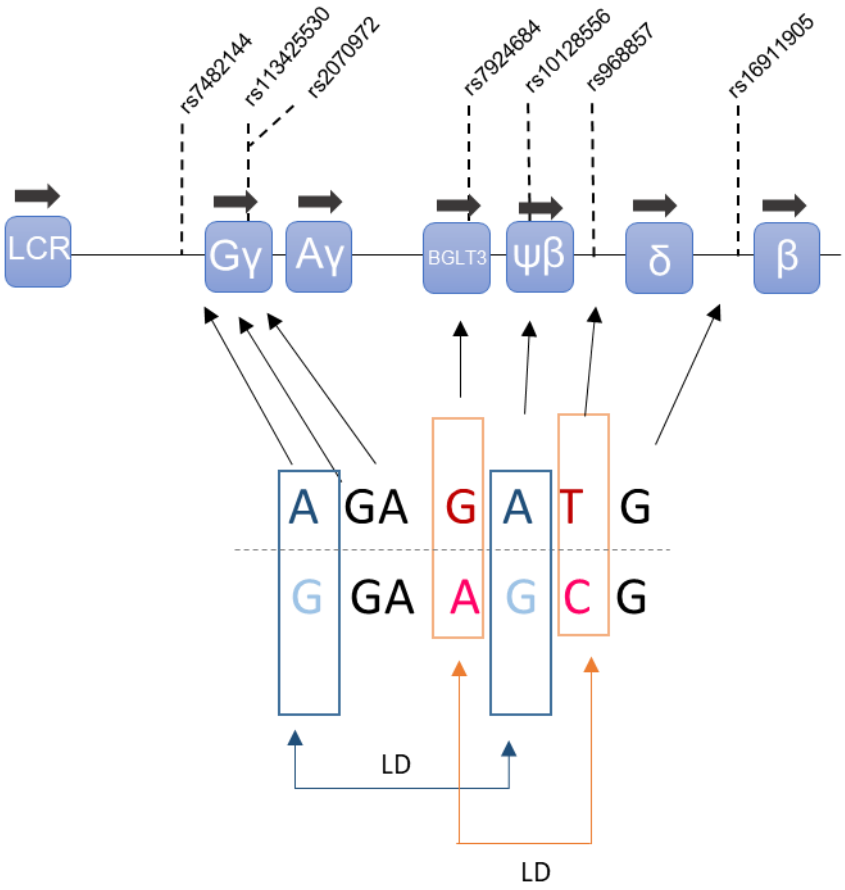


Figure 15. Schematic representation of the β-globin cluster containing the SNPs in this study and LD relationships between rs7482144 and rs101288556; and rs7924684 and rs968857.

This result suggests, on the one hand, that the combination of specific sequence features in the *HBG2* promoter and *HBBP1*, and on the other hand, the combination of features in *BGLT3* and the intergenic region between *HBBP1* and *HBD*, are crucial to modulate the levels of HbF. By other words, more important than the specific allele present at a given position, is the combination with alleles in other positions, producing long-range sequences that only in its whole can modulate levels of HbF.

5. Final Remarks

The *BGLT3* transcript and its transcription itself have been described as positive regulators of γ-globin region expression, even though being implied in distinct looping interactions (Ivaldi *et al.* 2018). However, for the first time, in this study, a variant present in this gene (rs7924684) was proven to be associated with high levels of HbF, in a sample of carriers of β-thalassemia mutations. The identified variant in *BGLT3* was found to be in very high LD with another, rs968857, located in the intergenic region between *HBBP1* and *HBD*, that was also strongly associated with high HbF levels.

Furthermore, two additional SNPs, one residing in the promoter of *HBG2* (rs7482144) and the other in the *HBBP1* pseudogene (rs10128556), were as well correlated with increased levels of HbF. These two SNPs were in very strong LD with each other, but not with the two SNPs previously mentioned.

Importantly, the associations here detected were much stronger than reported for any of the few known QTLs of fetal hemoglobin located outside the β-globin cluster, indicating that sequence variation throughout the entire β cluster plays the major role in HbF variability.

Taking in account our results, it seems that within the β-globin cluster, more important than the specific allele present at a given position is the combination of alleles at different SNPs, producing long-range sequences, that only in its whole can modulate levels of HbF.

Given that the 4 SNPs found associated with HbF were located in non-coding regions of the β-globin cluster, namely in the promoter of *HBG2*, in *BGLT3*, in the pseudogene *HBPP1* and the intergenic region between *HBBP1* and *HBD*, further deep studies are needed to understand how these 4 regions interact to modulate HbF.

In the context of therapeutic approaches aimed at HbF reactivation or induction, our data indicates that, at least in respect to the β-globin cluster, the targets must be haplotypes instead of the specific alleles in different SNPs.

6. Bibliographic References

Adelvand P.; Hamid M.; Sardari S. (2017). The Intrinsic Genetic and Epigenetic Regulator Factors as Therapeutic Targets, and the Effect on Fetal Globin Gene Expression. *Expert Review of Hematology.* 11(1):71-81.

Akinsheye I.; Alsultan A.; Solovieff N.; Ngo D.; Baldwin C.T.; Sebastiani P.; Chui D.H.K.; Steinberg M.H. (2011). Fetal hemoglobin in sickle cell anemia. *Blood.* 118(1):19-27.

Alhashem Y.N.; Vinjamur D.S.; Basu M.; Klingmüller U.; Gaensler K.M.; Lloyd J.A. (2011). Transcription factors KLF1 and KLF2 positively regulate embryonic and fetal beta-globin genes through direct promoter binding. *Journal of Biological Chemistry.* 286(28):24819-24827.

Almeida L.O. (2015). Human Gene Mutations and Migratory Flows – Portugal and the Mediterranean. *Advances in Anthropology.* 5:164.

Amaya M.; Desai M.; Gnanapragasam M.N.; Wang S.Z.; Zu Zhu S.; Williams D.C. Jr; Ginder G.D. (2013). Mi2β-mediated silencing of the fetal γ-globin gene in adult erythroid cells. *Blood.* 121(17):3493-3501.

Andrews N.C. (1998). The NF-E2 transcription factor. *International Journal of Biochemistry & Cell Biology.* 30(4):429-432.

Antoniani C.; Meneghini V.; Lattanzi A.; Felix T.; Romano O.; Magrin E.; Weber L.; Pavani G.; Hoss S.E.; Kurita R.; Nakamura Y.; Cradick T.J.; Lundberg A.S.; Porteus M.; Amendola M.; Nemer W.E.; Cavazzana M.; Mavilio F.; Miccio A. (2018). Induction of fetal hemoglobin synthesis by CRISPR/Cas9-mediated editing of the human β-globin locus. *Blood.* 131(17):1960-1973.

Armstrong R.A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics.* 34(5):502-508.

Barrett J.C.; Fry B.; Maller J.; Daly M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 21(2):263-265.

Bhatnagar P.; Purvis S.; Barron-Casella E.; DeBaun M.R.; Casella J.F.; Arking D.E.; Keefer J.R. (2011). Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *Journal of Human Genetics*. 56(4):316-323.

Bitoungui V.J.; Pule G.D.; Hanchard N.; Ngogang J.; Wonkam A. (2015). Beta-globin gene haplotypes among Cameroonians and review of the global distribution: is there a case for a single sickle mutation origin in Africa? *OMICS.* 19(3):171-179.

Blobel G.A.; Crossley M. (2018). Charting a noncoding gene for γ-globin activation. *Blood.* 132(18):1865-1867.

Borg J.; Papadopoulos P.; Georgitsi M.; Gutiérrez L.; Grech G.; Fanis P.; Phylactides M.; Verkerk A.J.; van der Spek P.J.; Scerri C.A.; Cassar W.; Galdies R.; van Ijcken W.; Ozgür Z.; Gillemans N.; Hou J.; Bugeja M.; Grosveld F.G.; von Lindern M.; Felice A.E.; Patrinos G.P.; Philipsen S. (2010). Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nature Genetics.* 42(9):801-805.

Brittain T. (2002). Molecular aspects of embryonic hemoglobin function. *Molecular Aspects of Medicine.* 23(4):293-342.

Buckingham L.; Flaws M.L. (2007). Molecular Diagnostics: Fundamentals, Methods & Clinical Applications. First edition. F. A. Davis.

Cabeda J.M.; Correia C.; Estevinho A.; Simões C.; Amorim M.L.; Pinho L.; Justiça B. (1999). Unexpected pattern of β-globin mutations in β-thalassemia patients from northern Portugal. *British Journal of Haematology*. 105:68-74.

Cao A.; Gossens M.; Piratsu M. (1989). B-thalassemia mutations in Mediterranean populations. *British Journal of Haematology.* 71:309-312.

Carter D.; Chakalova L.; Osborne C.S.; Dai Y.F.; Fraser P. (2002). Long-range chromatin regulatory interactions in vivo. *Nature Genetics.* 32:623-626.

Catani L.; Vianelli N.; Amabile M.; Pattacini L.; Valdrè L.; Fagioli M.E.; Poli M.; Gugliotta L.; Moi P.; Marini M.G.; Martinelli G.; Tura S.; Baccarani M. (2002). Nuclear factor-erythroid 2 (NF-E2) expression in normal and malignant megakaryocytopoiesis. *Leukemia.* 16(9):1773-1781.

Cavazzana M.; Antoniani C.; Miccio A. (2017). Gene Therapy for β-Hemoglobinopathies. *Molecular Therapy*. 25(5):1142-1154.

Cho Y.; Song S.H.; Lee J.J.; Kim C.G.; Dean A.; Kim A. (2008). The role of transcriptional activator GATA-1 at human beta-globin HS2. *Nucleic Acids Research.* 36(14):4521-4528.

Coutinho-Gomes M.P.; Costa M.G.G.; Braga L.B.; Cordeiro-Ferreira N.T.; Loi A.; Pirastu M.; Cao A. (1988). β-Thalassemia mutations in the portuguese population. *Human Genetics*. 78:13-15.

Couto C. (2017). Screening of *HBB* gene mutations in population samples from Alentejo and implementation of a SNaPshot® based system for HBB*S haplotyping (Master's thesis. Retrieved from the Open Repository of Universidade do Porto.

Crawford D.C.; Caggana M.; Harris K.B.; Lorey F.; Nash C.; Pass K.A.; Tempelis C.; Olney R.S. (2002). Characterization of β-globin haplotypes using blood spots from a population-based cohort of newborns with homozygous HbS. *Genetics in Medicine*. 4(5):328-335.

Crawford D.C.; Nickerson D.A. (2005). Definition and Clinical Importance of Haplotypes. *Annual Review of Medicine.* 56:303-320.

Das S.S.; Mitra A.; Chakravorty N. (2019). Diseases and their clinical heterogeneity – Are we ignoring the SNiPers and micRomaNAgers? An illustration using Beta-thalassemia clinical spectrum and fetal hemoglobin levels. *Genomics.* 111(1):67-75.

Dimovski A.J.; Oner C.; Agarwal S.; Gu Y.C.; Gu L.H.; Kutlar F.; Lanclos K.D.; Huisman T.H.J. (1991). Certain mutations observed in the 5' sequences of the $^{G}\gamma$- and $^{A}\gamma$-globin genes of β^{S} chromosomes are specific for chromosomes with major haplotypes. *Acta Hematologica.* 85(2):79-87.

Embury S.H.; Hebbel R.P.; Mohandas N.; Steinberg M.H. (1994). Sickle Cell Disease: basic principles and clinical practice. New York: Raven Press.

Excoffier L.; Lischer H.E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources.* 10:564-567.

Faustino P.; Osório-Almeida L.; Barbot J.; Espírito-Santo D.; Gonçalves J.; Romão L.; Martins M.C.; Marques M.M.; Lavinha J. (1992). Novel promoter and splice junctional defects add to the genetic, clinical or geographical heterogeneity of β-thalassemia in the Portuguese population. *Human Genetics.* 89:573-576.

Ferrone F.A. (2016). Sickle cell disease: its molecular mechanism and the one drug that treats it. *International Journal of Biological Macromolecules.* 93:1168-1173.

Forget B.C.; Bunn H.F. (2013). Classification of the Disorders of Hemoglobin. *Cold Spring Harbor Perspectives in Medicine*. 3(2): a011684.

Galanello R.; Cao A. (1998). Relationship between Genotype and Phenotype. Thalassemia Intermedia. *Annals New York Academy of Sciences.* 850:325-333.

Galanello R.; Origa R. (2010) Beta-thalassemia. *Orphanet Journal of Rare Diseases.* 5:11.

Galarneau G.; Palmer C.D.; Sankaran V.G.; Orkin S.H.; Hirschhorn J.N.; Lettre G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nature Genetics*. 42(12):1049-1051.

Giannopoulou E.; Bartsakoulia M.; Tafrali C.; Kourakli A.; Poulas K.; Stavrou E.F.; Papachatzopoulou A.; Georgitsi M.; Patrinos G.P. (2012). A single nucleotide polymorphism in the *HBBP1* gene in the human β-globin locus is associated with a mild β-thalassemia disease phenotype. *Hemoglobin*. 36(5):433-445.

Giardine B.; van Baal S.; Kaimakis P.; Riemer C.; Miller W.; Samara M.; Kollia P.; Anagnou N.P.; Chui D.H.; Wajcman H.; Hardison R.C.; Patrinos G.P. (2007). HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Human Mutation.* 28(2):206.

Guilman J.G.; Huisman T.H.J. (1985). DNA sequence variation associated with elevated fetal Gγ-globin production. *Blood.* 66(4):783-787.

Habara A.H.; Shaikho E.M.; Steinberg M.H. (2017). Fetal hemoglobin in sickle cell anemia: The Arab-Indian haplotype and new therapeutic agents. *American Journal of Hematology*. 92(11):1233-1242.

Hammer U.; Wegener R.; Nizze H.; Wöhlke G.; Kruse C.; Dworniczak B.; Kühn-Velten W.N.; Nöldge-Schomburg G.; Hofmockel R.; Jonas L. (2006). Sickle cell anemia: conclusions from a forensic case report of a young African woman who died after anesthesia. *Ultrastructural Pathology*. 30(6):415-422.

Hardison R.C.; Chui D.H.; Giardine B.; Riemer C.; Patrinos G.P.; Anagnou N.; Miller W.; Wajcman H. (2002). HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. *Human Mutation.* 19(3):225-233.

Harju-Baker S.; Costa F.C.; Fedosyuk H.; Neades R.; Peterson K.R. (2008). Silencing of Aγ-Globin Gene Expression during Adult Definitive Erythropioesis Mediated by GATA-1-FOG-1-Mi2 Complex Binding at the -566 GATA Site. *Molecular and Cellular Biology.* 28(10):3101-3113.

Hirokawa K.; Ohene-Frempong K.; Horiushi K. (1995). Determination of HbF level, maturation and morphology of individual sickle cell image cytometry. *Blood.* 86(10):139.

Honig G.R.; Adams J.G. (1986). Human Hemoglobin Genetics, First edition (Massachusetts: Blackwell).

Huang P.; Keller C.A.; Giardine B.; Grevet J.D.; Davies J.O.J.; Hughes J.R.; Kurita R.; Nakamura Y.; Hardison R.; Blobel G.A. (2017). Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. *Genes and Development*. 31(16):1704-1713.

Ivaldi M.S.; Diaz L.F.; Chakalova L.; Lee J.; Krivega I.; Dean A. (2018). Fetal γ-globin genes are regulated by the *BGLT3* long non-coding RNA locus. *Blood.* 132(18):1963-1973.

Jackson B.B.; Oruc A.C.; Kutlar A. (1996). A rapid PCR based non-radioactive method for the determination of β^S haplotypes. Abstracts of Papers, P33, 21st Annual Meeting of the National Sickle Cell Disease Program, Mobille, Alabama.

Keifer C.M.; Lee J.; Hou C.; Dale R.K.; Lee Y.T.; Meier E.R.; Miller J.L.; Dean A. (2011). Distinct Ldb1/NLI complexes orchestrate γ-globin repression and reactivation through ETO2 in human adult erythroid cells. *Blood*. 118(23):6200-6208.

Kent W.J. (2002). BLAT – the BLAST-like alignment tool. *Genome Research.* 12:656-664.

Kibbe W.A. (2007). OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Research.* 35: W43-W46.

Kulozik A.E.; Yarwood N.; Jones R.W. (1988). The Corfu delta beta zero thalassemia: a small deletion acts at a distance to selectively abolish beta globin gene expression. *Blood.* 71(2):457-462.

Labie D.; Pagnier J.; Lapoumeroulie C.; Rouabhi F.; Dunda-Belkhodja O.; Chardin P.; Beldjord C.; Wajcman H.; Fabry M.E.; Nagel R.L. (1985). Common haplotype dependency of high ^{G}γ-globin gene expression and high Hb F levels in β-thalassemia and sickle cell anemia patients. *Proceedings of the National Academy of Sciences of the United States of America.* 82(7):2111-2114.

Lanclos K.D.; Oner C.; Dimovski A.J.; Gu Y.C.; Huisman T.H.J. (1991). Sequence variations in the 5' flanking and IVS-II regions of the ^{G}γ- and ^{A}γ-globin genes of β^S chromosomes with five different haplotypes. *Blood.* 77(11):2488-2496.

Lavinha J.; Baiget M. (1992). Beta thalassemia in Spain and Portugal: epidemiology and molecular pathology. *Hematolgy Reviews.* 6:113-116.

Leal A.S.; Martins P.R.J.; Balarin M.A.S.; Pereira G.A.; Resende G.A.D. (2016). Haplotypes βs-globin and its clinical-haematological correlation in patients with sickle-cell anemia in Triângulo Mineiro, Minas Gerais, Brazil. *Jornal Brasileiro de Patologia e Medicina Laboratorial.* 52(1):6-10.

Lettre G.; Bauer D.E. (2016). Fetal haemoglobin in sickle cell disease: from genetic epidemiology to new therapeutic strategies. *Lancet.* 387:2554-2564.

Liu L.; Pertsemlidis A.; Ding L.H.; Story M.D.; Steinberg M.H.; Sebastiani P.; Hoppe C.; Ballas S.K.; Pace B.S. (2016) A case-control genome-wide association study identifies genetic modifiers of fetal hemoglobin in sickle cell disease. *Experimental Biology and Medicine*. 241(7):706-718.

Mandal P.K.; Kartthik S. (2019) Autoimmune hemolytic anemia: an uncommon cause of elevated fetal hemoglobin. *Journal of Hematopathology*.

Marengo-Rowe A.J. (2006). Structure-function relations of human hemoglobins. *Proceedings (Baylor University. Medical Center)*. 19(3):239-245.

Martins M.C.; Olim G.; Melo J.; Magalhães H.Á.; Rodrigues M.O. (1993). Hereditary anemias in Portugal: epidemiology, public health significance and control. *Journal of Medical Genetics.* 30:235-239.

Martyn G.E.; Weinert B.; Yang L.; Shah M.; Norton L.J.; Burdach J.; Kurita R.; Nakamura Y.; Pearson R.C.M.; Funnell A.P.W.; Quinlan K.G.R.; Crossley M. (2018). Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding. *Nature Genetics*. 50(2018):498-503.

Masuda T.; Wang X.; Maeda M.; Canver M.C.; Sher F.; Funnell A.P.; Fisher C.; Suciu M.; Martyn G.E.; Norton L.J.; Zhu C.; Kurita R.; Nakamura Y.; Xu J.; Higgs D.R.; Crossley M.; Bauer D.E.; Orkin S.H.; Kharchenko P.V.; Maeda T. (2016). Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. *Science.* 351(6270):285-289.

Menzel S.; Thein S.L. (2018). Genetic Modifiers of Fetal Hemoglobin in Sickle Cell Disease. *Molecular Diagnosis & Therapy*. 23(2):235-244.

Michlitsch J.G.; Walters M.C. (2008). Recent advances in bone marrow transplantation in hemoglobinopathies. *Current Molecular Medicine.* 8:675-689.

Modell B.; Darlison M. (2008). Global epidemiology of haemoglobin disorders and derived service indicators. *Bulletin of the World Health Organization.* 86(6):480-487.

Moleirinho A.; Seixas S. Lopes A.M.; Bento C.; Prata M.J.; Amorim A. (2013). Evolutionary constraints in the β-globin cluster: the signature of purifying selection at the δ-globin (HBD) locus and its role in developmental gene regulation. *Genome Biology and Evolution.* 5(3):559-571.

Mozeleski B.M.; Al-Rubaish A.; Al-Ali A.; Romero J. (2018). Perspective: A Novel Prognostic for Sickle Cell Disease. *Saudi Journal of Medicine & Medical Sciences*. 6(3):133-136.

Mpollo M.S.; Beaudoin M.; Berg P.E.; Beauchemin H.; D'Agati V.; Trudel M. (2006). BP1 is a negative modulator of definitive erythropoiesis. *Nucleic Acids Research.* 34(18):5232-5237.

Nuinoon M.; Makarasara W.; Mushiroda T *et al.* (2010). A genome-wide association identified the common genetic variants influence disease severity in β^0-thalassemia/hemoglobin E.

Orkin S.H.; Kazazian Jr. H.H.; Antonarakis S.E.; Goff S.C.; Boehm C.D.; Sexton J.P.; Waber P.G.; Giardina P.J. (1982). Linkage of β-thalassemia mutations and β-globin gene polymorphisms with DNA polymorphisms in human β-globin gene cluster. *Nature.* 296(5858):627-631.

Orkin S.H.; Bauer D.E. (2019). Emerging Genetic Therapy for Sickle Cell Disease. *Annual Review of Medicine*. 70:257-271.

Palstra R.J.; Tolhius B.; Splinter E.; Nijmeijer R.; Grosveld F.; de Laat W. (2003). The β-globin nuclear compartment in the development and erythroid differentiation. *Nature Genetics.* 35:190-194.

Patrinos G.P.; Giardine B.; Riemer C.; Miller W.; Chui D.H.K.; Anagnou N.P.; Wajcman H.; Hardison R.C. (2004). Improvements in the *HbVar* database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. *Nucleic Acid Research*. 32(Database issue): D537-D541.

Pereira C.; Relvas L.; Bento C.; Abade A.; Ribeiro M.L.; Manco L. (2015). Polymorphic variations influencing fetal hemoglobin levels: Association study in beta-thalasemia

carriers and in normal individuals of Portuguese origin. *Blood Cells, Molecules and Diseases*. 54:315-320.

Persons D.A. (2009). Hematopoietic stem cell gene transfer for the treatment of hemoglobin disorders. *Hematology.* 2009:690-697.

Philipsen S.; Hardison R.C. (2018) Evolution of hemoglobin loci and their regulatory elements. *Blood Cells, Molecules and Diseases.* 70(2018):2-12.

Powars D.R. (1991). Sickle cell anemia: βS-gene-cluster haplotypes as prognostic indicators of vital organ failure. *Seminars in Hematology*. 28(3):202-208.

Purcell S.; Neale B.; Todd-Brown K.; Thomas L.; Ferreira M.A.; Bender D.; Maller J.; Sklar P.; de Bakker P.I.; Daly M.J.; Sham P.C. (2007). PLINK: a tool set for whole genome association and population-based linkage analyses. *American Journal of Human Genetics.* 81(3):559-575.

Roseff S.D. (2009). Sickle cell disease: a review. *Immunohematology.* 25(2):67-74.

Sachidanandam R. *et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature.* 409(6822):928-933.

Sankaran V.G.; Menne T.F.; Xu J.; Akie T.E.; Lettre G.; Van Handel B.; Mikkola H.K.; Hirschhorn J.N.; Cantor A.B.; Orkin S.H. (2008). Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science.* 322(5909):1839-1842.

Sankaran V.G.; Xu J.; Orkin S.H. (2010). Advances in the understanding of haemoglobin switching. *British Journal of Haematology.* 149:181-194.

Sankaran V.G.; Xu J.; Byron R.; Greisman H.A.; Fisher C.; Weatherall D.J.; Sabath D.E.; Groudine M.; Orkin S.H.; Permawardhena A.; Bender M.A. (2011). A functional element necessary for fetal hemoglobin silencing. *The New England Journal of Medicine.* 365(9):807-814.

Sankaran V.G.; Orkin S.H. (2013). The Switch from Fetal to Adult Hemoglobin. *Cold Spring Harbor Perspectives in Medicine.* 3(1): a011643.

Sankaran V.G.; Weiss M.J. (2015). Anemia: progress in molecular mechanisms and therapies. *Nature Medicine.* 21:221-230.

Shang X.; Xu X. (2017). Update in the genetics of thalassemia: What clinicians need to know. *Best Practice & Research Clinical Obstetrics and Gynaecology.* 39(3):3-15.

Sherwood L. (2015). Human Physiology: From Cells to Systems, Ninth edition. (Boston: Cengage Learning).

Shriner D.; Rotimi C.N. (2018). Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sickle Allele during the Holocene Wet Phase. *American Journal of Human Genetics.* 102(4):547-556.

Stadhouders R.; Aktuna S.; Thongjuea S.; Aghajanirefah A.; Pourfarzad F.; van Ijcken W.; Lenhard B.; Rooks H.; Best S.; Menzel S.; Grosveld F.; Thein S.L.; Soler E. (2014). HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *Journal of Clinical Investigation.* 124(4):1699-1710.

Steinberg M.H.; B.G. Forget; Higgs D.R.; Weatherall D.J. (2001). Disorders of Hemoglobin: Genetics, Pathophysiology and Clinical Management. *Journal of the Royal Society of Medicine.* 94(11):602-603.

Steinberg M.H. (2009). Genetic etiologies for phenotypic diversity in sickle cell anemia. *The Scientific World Journal*. 9:46-67.

Steinberg M.H.; McCarthy W.F.; Castro O.; Ballas S.K.; Armstrong F.D.; Smith W.; Ataga K.; Swerdlow P.; Kutlar A.; DeCastro L.; Waclawiw M.A. (2010). The risks and benefits of long-term use of hydroxyurea in sickle cell anemia: a 17.5-year follow-up. *American Journal of Hematology*. 85:403-408.

Steinberg M.H.; Sebastiani P. (2012). Genetic modifiers of sickle cell disease. *American Journal of Hematology*. 87:795-803.

Steinberg M.H.; Chui D.H.K.; Dover G.J.; Sebastiani P.; Alsultan A. (2014). Fetal hemoglobin in sickle cell anemia: a glass half full? *Blood.* 123:481-485.

Storz J.F. (2016). Gene Duplication and Evolutionary Innovations in Hemoglobin-Oxygen Transport. *Physiology.* 31(3):223-232.

Sutton M.; Bouhassira E.; Nagel R.L. (1989). Polymerase chain reaction amplification applied to the determination of the β-like globin gene cluster haplotypes. *American Journal of Hematology.* 32(1):66-69.

Syvänen A.C. (2001). Acessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*. 2(12):930-942.

Tallack M.R.; Perkins A.C. (2013). Three fingers on the switch: Krüppel-like factor 1 regulation of γ-globin to β-globin gene switching. *Current Opinion in Hematology.* 20(3):193-200.

Tamagnini G.P.; Lopes M.C.; Castanheira M.E.; Wainscoat J.S. (1983). β^+ Thalassemia – Portuguese type: clinical, haematological and molecular studies of a newly defined form of β thalassemia. *British Journal of Haematology.* 54:189-200.

Thein S.L.; Menzel S.; Lathrop M.; Garner C. (2009). Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. *Human Molecular Genetics.* 18(R2): R216-R223.

Thom C.S.; Dickson C.F.; Gell D.A.; Weiss M.J. (2013). Hemoglobin Variants: Biochemical Properties and Clinical Correlates. *Cold Spring Harbor Perspectives in Medicine*. 3(3):a011858.

Thomas C.; Lumb A. (2012). Physiology of haemoglobin. *Continuing Education in Anaesthesia Critical Care & Pain*. 12(5):251-256.

Tolhuis B.; Palstra R.J.; Splinter E.; Grosveld F.; de Laat W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell.* 10:1453-1465.

Traxler E.A.; Yao Y.; Wang Y.D.; Woodard K.J.; Kurita R.; Nakamura Y.; Hughes J.R.; Hardison R.C.; Blobel G.A.; Li C.; Weiss M.J. (2016). A genome-editing strategy to treat β-hemoglobinopathies that recapitulates a mutation associated with a benign genetic condition. *Nature Medicine.* 22:987-990.

Untergasser A.; Cutcutache I.; Koressaar T.; Ye J.; Faircloth B.C.; Remm M.; Rozen S.G. (2012). Primer3 – new capabilities and interfaces. *Nucleic Acids Research.* 40: e115.

Venter J.C. *et al.* (2001). The sequence of the human genome. *Science.* 291(5507):1304-1351.

Vinson A.E.; Walker A.; Elam D.; Glendenning M.; Kutlar F.; Clair B.; Harbin J.; Kutlar A. (2004). A Novel Approach to Rapid Determination of β^S-Globin Haplotypes: Sequencing of the $^A\gamma$-IVS-II Region. *Hemoglobin.* 28(4):317-323.

Voskaridou E.; Christoulas D.; Bilalis A.; Plata E.; Varvagiannis K.; Stamatopoulos G.; Sinopoulou K.; Balassopoulou A.; Loukopoulos D.; Terpos E. (2010). The effect of prolonged administration of hydroxyurea on morbidity and mortality in adult patients with sickle cell syndromes: results of a 17-year, single-center trial (LaSHS). *Blood.* 115:2354-2363.

Wall J.D.; Pritchard J.K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics.* 4(8):587-597.

Wang X.; Thein S.L. (2018). Switching from fetal to adult hemoglobin. *Nature Genetics.* 50(2018):478-480.

Weatherall D.J. (1986). The thalassemias: molecular pathogenesis. In: Bunn H.T. & Forget B.G.: Hemoglobin: Molecular, Genetics and Clinical Aspects. *W.B. Saunders.* pp. 213-321.

Weatherall D.; Akinyanju O.; Fucharoen S.; Olivieri N.; Musgrove P. (2006). Inherited Disorders of Hemoglobin. In: Jamison D.T.; Breman J.G.; Measham A.R.; *et al*. editors. Disease control priorities in developing countries. *Oxford University Press*. pp:119-138.

Wienert B.; Martyn G.E.; Funnell A.P.W.; Quinlan K.G.R.; Crossley M. (2018) Wake-up Sleepy Gene: Reactivating Fetal Globin for β-Hemoglobinopathies. *Trends in Genetics.* 34(12):927-940.

Wong T.E.; Brandow A.M.; Lim W.; Lottenberg R. (2014). Update on the use of hydroxyurea therapy in sickle cell disease. *Blood.* 124:3850-3857.

Woon Kim Y.; Kim S.; Geun Kim C.; Kim A. (2011). The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal γ-globin genes. *Nucleic Acids Research.* 39(16):6944-6955.

Xu J.; Sankaran V.G.; Ni M.; Menne T.F.; Puram R.V.; Kim W.; Orkin S.H. (2010). Transcriptional silencing of γ-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes & Development.* 24(8):783-798.

Yates A.; Akanni W.; Amode M.R.; Barrel D.; Billis K.; Carvalho-Silva D.; Cummins C.; Clapham P.; Fitzgerald S.; Gil L. (2015). Ensembl 2016. *Nucleic Acids Research.* 44: D710-D716.

Zhou D.; Pawlik K.M.; Ren J.; Sun C.W.; Townes T.M. (2006). Differential binding of erythroid Krupple-like factor to embryonic/fetal globin gene promoters during development. *Journal of Biological Chemistry.* 281(23):16052-16057.

Zhou D.; Liu K.; Sun C.W.; Pawlik K.M.; Townes T.M. (2010). KLF1 regulates BCL11A expression and gamma-to-beta-globin gene switching. *Nature Genetics.* 42(9):742-744.

Supplementary Data

Sample	rs7482144	rs113425530	rs2070972	rs7924684	rs10128556	rs968857	rs16911905
28	A A	G G	A A	G G	A A	T T	G G
86	G G	G G	A A	G A	G G	C T	G G
127	G A	G G	A C	G A	G A	C T	G G
130	A A	G G	A A	G G	A A	T T	G G
132	G A	G G	A C	G A	G A	C T	G G
142	G G	G G	A C	G A	G G	C T	G G
155	G G	G G	A A	G A	G G	C T	G G
172	G G	G G	A C	G A	G G	C T	G G
236	G A	G G	A C	G A	G A	C T	G G
(3356) 238	G A	G G	A C	G A	G A	C T	G G
250	G G	G G	A C	A A	G G	C C	G G
283	G G	G G	A C	G A	G G	C T	G G
(2655) 341	G G	G G	A C	G A	G G	C T	G C
385	G G	G G	A C	A A	G G	C C	G G
397	G A	G G	A A	G G	G A	T T	G G
402	G G	G G	C C	A A	G G	C C	G G
406	G G	G G	A C	G A	G G	C C	G G
427	G G	G G	A C	A A	G G	C C	G G
(2644) 435	G G	G G	A C	G A	G G	C C	G C
458	G G	G G	C C	A A	G G	C C	G G
463	G G	G G	A C	G A	G G	C C	G G
512	G A	G G	A C	G A	G A	C T	G G
538	G A	G G	A C	G A	G A	C T	G G
539	G G	G G	C C	A A	G G	C C	G G
542	G A	G G	A A	G A	G A	C T	G G
577	G A	G G	A A	G A	G A	C T	G G
(3310) 586	G G	G G	A C	G A	G G	C C	G G
630	G A	G G	A C	G A	G A	C C	G G
638	G G	G G	A C	G A	G G	C T	G G
692	G G	G G	A C	G A	G G	C C	G G
697	G G	G G	A C	G A	G G	C T	G C
726	G G	G G	A C	G A	G A	C T	G G
751	G G	G G	A C	A A	G G	C T	G G
799	G G	G G	A A	G G	G G	C T	G C
803	G G	G T	A A	G A	G G	C T	G G
807	G G	G G	A C	A A	G G	C C	G G
880	G G	G G	A C	A A	G G	C C	G G
892	A A	G G	A A	G G	A A	T T	G C

896	G G	G G	A C	G A	G G	C T	G G
900	G G	G G	A A	A A	G G	C C	G G
902	G A	G G	A A	G G	G A	T T	G G
928	G A	G G	A A	G G	G A	T T	G G
976	G G	G G	A C	A A	G G	C C	G G
1012	G G	G G	A A	G G	G A	T T	G G
1092	G A	G G	A A	G G	G A	T T	G G
1603	G A	G G	A C	G A	G A	C T	G G
1687	G G	G G	A C	A A	G G	C C	G G
1714	G A	G G	A A	G G	G A	T T	G C
1762	G G	G G	A C	G A	G G	C T	G G
1764	G A	G G	A C	G A	G A	C T	G G
1806	G G	G G	A C	G A	G G	C T	G G
1817	G G	G G	A A	G A	G G	C T	G G
2040	G G	G G	A C	G A	G G	C T	G G
2096	G A	G G	A C	G A	G A	C T	G G
2871	G G	G G	A C	A A	G G	C C	G G
3810	G G	G G	A C	G A	G G	C T	G C
4343	G G	G G	A A	G G	G G	T T	G C
4860	G G	G G	A A	G A	G G	C T	G G
4938	G G	G G	A C	G A	G G	C T	G C
5015	G G	G G	A C	A A	G G	C C	G G
5810	G G	G G	A C	A A	G G	C C	G C
6629	G G	G G	A C	A A	G G	C C	G G
6666	G A	G G	A A	G G	G A	T T	G C
7053	G G	G G	A A	G G	G G	T T	G G
7096	G G	G G	A A	A A	G G	C C	G G
7189	G G	G G	A C	G A	G G	C T	G G
8484	G G	G G	A C	G A	G G	C T	G C
10132	G G	G G	A C	A A	G G	C C	G G
24403	G A	G G	A A	G G	G A	T T	G C
63262	G G	G G	A A	A A	G G	C C	G G
91159	A A	G G	A A	G G	A A	T T	G G

Table S1. Genotypic data of the polymorphisms found in the 71 studied samples from Portuguese individuals with β-thalassemia.