Assessing the role of genetic variations at the β-globin gene cluster in levels of fetal hemoglobin

Catarina Rocha Pacheco

Mestrado em Genética Forense

Departamento de Biologia

2020

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).

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

30 ANOS
A DESCOBRIR O CANCRO
ipatimup

INSTITUTO
DE INVESTIGAÇÃO
E INOVAÇÃO
EM SAÚDE
UNIVERSIDADE
DO PORTO

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.
O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

A toda a equipa do grupo de Genética Populacional e Evolução, do Instituto de Investigação e Inovação em Saúde (i3S), pela hospitalidade com que me receberam.

À minha orientadora Professora Doutora Maria João Prata, pela excelente orientação que me proporcionou, por todos os conhecimentos que me transmitiu, pela total dedicação, preocupação e compreensão. Pelo tempo que dispensou para me ajudar. Obrigada por todas as sugestões, incentivo e ajuda na elaboração desta dissertação. Foi sem dúvida um privilégio poder aprender consigo.

À minha coorientadora Doutora Verónica Gomes, por todo o auxílio e acompanhamento na parte laboratorial deste trabalho e pela incansável disponibilidade e paciência.

Ao Doutor Licínio Manco, pela amabilidade e disponibilidade prestada e cuja ajuda foi indispensável para a concretização deste trabalho.

Aos meus pais por todo o carinho e amor incondicional, por acreditarem em mim e me motivarem sempre a seguir os meus sonhos. Nunca poderei agradecer o suficiente o vosso apoio em todas as etapas da minha vida.

O meu sincero obrigada a todos.

Abstract

Fetal hemoglobin (HbF) is the predominant type of hemoglobin during fetal life. After birth, the synthesis of HbF decreases rapidly as it is replaced by adult forms of hemoglobin, so that in healthy adults it usually represents <1% of the total hemoglobin. However, adult inter-individual levels of HbF vary substantially, which is largely determined by genetic factors. High HbF levels during adulthood are considered to have little physiological and clinical significance in healthy individuals, but in β-hemoglobinopathies patients they are associated with an ameliorated clinical course of the diseases. Hence, therapeutic approaches aimed at stimulating the production of HbF are of greatly promising.

Even though common polymorphisms at the β-globin gene cluster have already been suggested as modifiers of HbF levels, the degree and condictions they contribute to HbF expression are still under discussion.

In this work, we took advantage of a 7-Plex SNaPshot® System in order to evaluate the effects of seven single nucleotide polymorphisms (SNPs) within the β-globin cluster in levels of HbF in a sample of 51 Portuguese individuals lacking any hematological disorder. The System includes 6 SNPs that define the conventional HBB*S haplotypes plus one SNP located in the BGLT3 gene that was recently found to be strongly associated with HbF levels in Portuguese β-thalassemia carriers.

Our results suggest that the SNPs at the β-globin clusters that more strongly were correlated with HbF levels in β-hemaglobinopathies patients, had little or null influence in normal subjects. It appeared, therefore, that under conditions of erythropoietic stress induced by disorders such β-thalassemia and probably SCD, regulatory mechanisms of HbF levels in which participate interactively a variety of sequences from the β- cluster might be activated, while its role in normal condictions is irrelevant or much more silent.

Haplotypic analysis also led to conclude that physiological conditions still poorly understood seem to be important to induce molecular strategies regulating levels of HbF that may imply different sequences in the β-cluster.

Keywords: Fetal hemoglobin (HbF), Single Nucleotide Polymorphisms (SNPs), β-globin cluster, SNaPshot® Multiplex, Haplotypes.

Resumo

A hemoglobina fetal (HbF), é o tipo de hemoglobina predominante durante o desenvolvimento fetal. No período pós-natal, a síntese de HbF diminui rapidamente à medida que é substituída pelas formas adultas de hemoglobina, de tal modo que em adultos saudáveis representa normalmente <1% da hemoglobina total. Contudo, existe grande variabilidade quanto a níveis de HbF entre indivíduos, sabendo-se que tal é parcialmente determinado por fatores geneticos. Níveis elevados de HbF durante a vida adulta têm pouca revelância fisiológica e clínica em indivíduos normais, no entanto, em doentes com β-hemoglobinopatias estão associados a manifestações clínicas menos severas. Deste modo, o desenvolvimento de novas terapias que visam o aumento de HbF tem assumido grande importância.

Ainda que alguns polimorfismos comuns no cluster das β-globinas tenham já sido descritos como modificadores genéticos dos níveis de HbF, o grau e condições em que contribuem para alterar níveis de HbF ainda não são claros

Neste trabalho, tiramos partido de um sistema 7-Plex SNaPshot® com o objetivo de avaliar os efeitos de sete SNPs dispersos pelo cluster das β-globinas nos níveis de HbF, analisando uma amostra de 51 indivíduos portugueses sem qualquer distúrbio genético do foro hematológico. O sistema inclui 6 SNPs que definem os haplótipos associados à mutação que determina anemia falciforme, bem como um SNP no gene BGLT3 que foi recentemente associado com variabilidade nos níveis de HbF em portadores de β-talassemia de origem portuguesa.

Os resultados obtidos revelaram que os SNPs do cluster das β-globinas que demonstraram fortes sinais de associação com níveis de HbF em pacientes com β-hemoglobinopatias, pouca influencia têm em indivíduos saudáveis. Assim, é de admitir que em condições de stress fisiológico resultante de patologias, como acontece nas β-talassemias e na anemia falciforme, sejam ativados mecanismos de regulação dos níveis de HbF em que participam várias sequências do cluster das β-globinas, cujo papel em condições normais parece ser muito menos relevante.

A análise dos haplótipos associados aos 7 SNPs também sugeririu que condições fisiológicas ainda não totalmente conhecidas parecem ser importantes para desencadear estratégias reguladoras dos níveis de HbF em que participam diferentes sequências do cluster HBB.

Palavras-chave: Hemoglobina fetal, Single Nucleotide Polymorphisms (SNPs), Cluster HBB, SNaPshot® Multiplex, Haplótipos.

Table of contents

List of Figures

List of Tables

List of Abbreviations

A	Adenine
α	Alpha
$\alpha_2\varepsilon_2$	Hemoglobin Gower 2
AI	Arab-Indian
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
C	Cytosine
CAM	Cameroon
CAR	Central African Republic
CI	Confidence Interval
CO_2	Carbon Dioxide
CRISPR-Cas9	Clustered Regularly Interspaced Short Palindromic Repeat-associated 9
◦C	Degree Celcius
δ	Delta
ddNTP	dideoxynucleotide
DNA	Deoxyribonucleic Acid
ε	Epsilon
EM/ELB	Expectation Maximization/Excoffier-Laval-Balding
ExoSAP	Exonuclease I Shrimp Alkaline Phosphatase
FastAP	Thermosensitive Alkaline Phosphatase
FW	Forward
G	Guanine
γ	Gamma
GWAS	Genome Wide Association Study
h	Hour
Hb	Hemoglobin

HbA ($\alpha_2\beta_2$)	Adult Hemoglobin - major form
HbA2 ($\alpha_2\delta_2$)	Adult Hemoglobin - minor form
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 ($\alpha_2\gamma_2$)	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
HSs	Hipersensitive Sites
HSCT	Hematopoietic Stem Cell Transplant
HWE	Hardy-Weinberg Equilibrium
∞	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	Micromolar
min	Minute
NF-E2	Nuclear Factor Erythroid 2
NuRD	Nucleossome Remodeling Deacetylase
O2	Oxygen
OR	Odds Ratio
ψ	Psi
%	Percentage
PCR	Polymerase Chain Reaction
®	Registered
RBCs	Red Blood Cells
RFLP	Restriction Fragment Length Polymorphism

RNA	Ribonucleic Acid
RNA Pol II	RNA polimerase II
rs	Reference SNP
RV	Reverse
SBE	Single Base Extension
SCD	Sickle Cell Disease
SD	standard Deviation
SE	Standard Error
sec	Second
SEN	Senegal
SNPs	Single Nucleotide Polymorphisms
θ	Theta
T	Thymine
TM	Trade Mark
V	Volt
ζ	Zeta
$\zeta_2\gamma_2$	Hemoglobin Portland
$\zeta_2\varepsilon_2$	Hemoglobin Gower 1
ZBTB7A	Zinc finger and BTB domain-containing protein 7A

1. Introduction

1.1. The Hemoglobin Protein

Hemoglobins (Hbs) are proteins found in red blood cells (RBCs), that give the blood its red shade. Their primary function is to carry oxygen (O_2) from the lungs to peripheral tissues, where O_2 is necessary for aerobic metabolism, and to help return to the lungs carbon dioxide (CO_2) to be expelled (Philipsen & Hardison, 2018).

All forms of hemoglobins are tetrameric proteins composed of two “α-like” and two “β-like” polypeptide chains that are similar in length but differ in amino acid sequences. Each chain contains a heme-iron complex, where O_2 reversibly binds (Figure 1A).

Depending on the developmental stage and in response to the variation of oxygen requirements, different types of “α-like” and “β-like” globin molecules are produced, leading to forms of Hb that are present at distinct times during development, i.e. embryonic, fetal, and adult stages. At the embryonic stage, the forms of embryonic hemoglobins that start being produced are: Hb Gower 1 ($\zeta_2\varepsilon_2$), Hb Gower 2 ($\alpha_2\varepsilon_2$) and Hb Portland ($\zeta_2\gamma_2$). At the fetal stage, RBCs contain fetal hemoglobin ($\alpha_2\gamma_2$) which decreases rapidly in the postnatal period as it is replaced by the adult forms of hemoglobin: HbA ($\alpha_2\beta_2$) and HbA2 ($\alpha_2\delta_2$) (Figure 1B) (Marengo-Rowe, 2006).

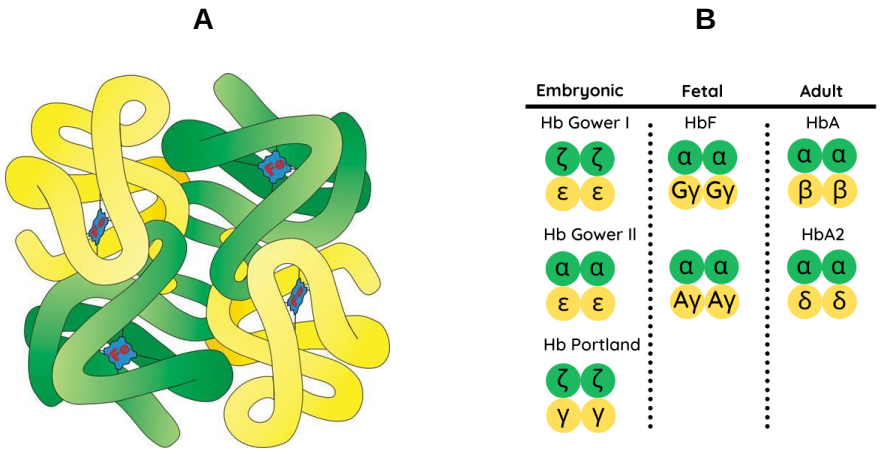


Figure 1 – The hemoglobin protein. **A**: Schematic representation of the structure of the hemoglobin molecule: two identical α-globin chains (green) and two identical β-globin chains (yellow). Each chain contains a heme-iron complex (blue/red) (Thomas & Lumb, 2012) **B**: Diffferent tetrameric combinations of the embryonic, fetal and adult forms of Hb.

The genes encoding the different globin chains belong to two clusters of genes. The α-globin cluster, located on chromosome 16, contains four pseudogenes: HBZP1 (ψζ), HBM (ψα2) HBAP1 (ψα1) and HBQ1 (θ1); and three genes: HBZ (ζ), and the duplicated HBA (α1) and HBA2 (α2), the last two codifying for α chains. The β-globin cluster, located on chromosome 11, comprises five genes coding for "β-like" globins: the embryonic-specific HBE1 (ε), the duplicated and homologous fetal-predominant HBG2 (Gγ) and HBG1 (Aγ), and the adult expressed HBD (δ) and HBB (β) globin genes. The intergenic region between the γ-globin and δ-globin genes contains the pseudogene HBBP1 (ψβ) and the BGLT3 gene that codes for an erythroid-specific long noncoding RNA (Figure 2) (Huang *et al*. 2017; Ivaldi *et al*. 2018; Waye & Chui, 2001).

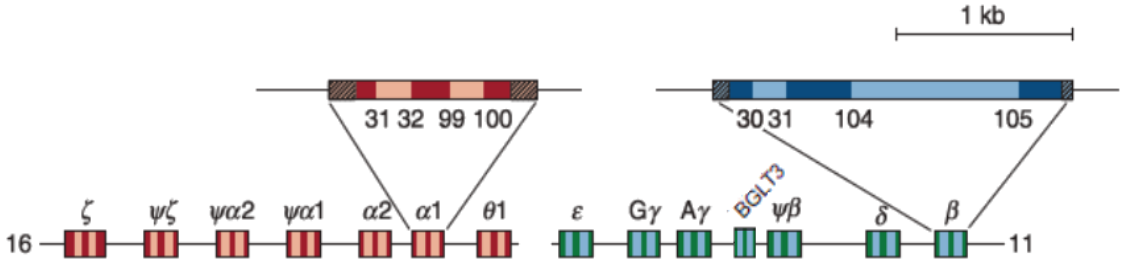


Figure 2 – Representation of the α-globin cluster (red) and β-globin cluster (blue) genes (Adapted from Couto, 2017).

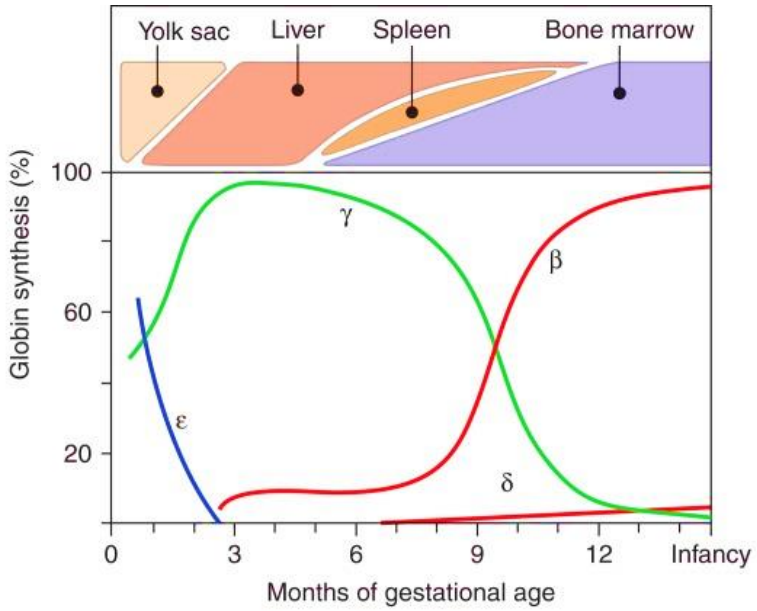
1.1.1. Globin Gene Switch

The first developmental switch in the expression of genes from the β-globin cluster occurs after the early part of the first trimester of pregnancy when the expression of ε-globin is silenced, and γ-globin genes are upregulated. Soon after the time of birth, it follows the second switch from predominant expression of fetal γ-globin to adult β-globin. The adult δ-globin also starts being produced but it is poorly expressed (Sankaran & Orkin, 2013; Cavazzana *et al*. 2017) (Figure 3A).

The locus control region (LCR) is a 16-kb-long cis-regulatory element required for normal expression of the "β-like" globin genes. It has a powerful role in chromatin-opening and contains five DNA enhancer elements (DNase hypersensitive sites [HSs - HS1, HS2, HS3, HS4 and HS5]), crucial for high-level globin gene expression (Figure 3B). The LCR loops to the γ–globin promoter in fetal erythroblasts, and to the β-globin promoter in adult erythroblasts to regulate gene expression.

Several transcription factors are implicated in the hemoglobin switching process, influencing "β-like" globin gene expression by interacting directly with genes from the β-globin cluster, or indirectly through association with other transcription factors. The erythroid Kruppel-like factor (KLF1), the stage selector protein MYB, the B-cell lymphoma/leukemia 11A (BCL11A), and the lymphoma/leukemia-related factor (LRF) are some of the known factors (Cavazzana *et al*. 2017). Their role in the hemoglobin switch will be further discussed later.

A



B

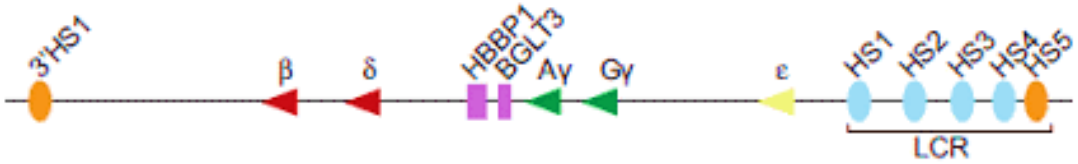


Figure 3 - Hemoglobin switch. **A**: During the embryonic stage, ε-globin is predominantly expressed, until it starts to be replaced by γ-globin during fetal life. Soon after birth and during adult life, the β-globin becomes the major "β-like" globin. A small amout of the adult δ-globin is also produced. (Sankaran & Orkin, 2013) **B**: Diagram of the β-globin cluster showing the locus control region (LCR) with its DNase hypersensitivity sites (HSs- HS1, HS2, HS3, HS4 and HS5) (Huang *et al*. 2017).

1.1.2. Inherited Hemoglobin Disorders

Hemoglobin disorders are the most common inherited blood disorders. They can be caused by impaired synthesis of the globin chains, with reduced or total absence of one of the globin subunits (thalassemia syndromes), or by a structural defect in one of the globin subunits, which creates different Hb tetramer "variants". These variants result mostly from missense mutations that cause single amino acid substitutions in the globin protein. Often, they are innocuous but in some cases they can result in a nonfunctional protein, leading to a clinical disorder (Forget & Bunn, 2013).

Sickle cell disease (SCD) and β-thalassemia are β-hemoglobinopathies which result from mutations in the HBB gene and are among the most common monogenetic diseases in the world (Cavazzana *et al*. 2017). These conditions contribute significantly to morbidity and mortality worldwide. Apart from rare cases, they both follow a recessive mode of inheritance with carriers or heterozygotes being usually symptomless.

SCD is caused by a single A-to-T point mutation at codon six of the β-globin gene (HBB), which replaces glutamic acid by valine forming hemoglobin S (HbS, $\alpha_2\beta^S{}_2$). HbS has the capacity to carry oxygen but in its deoxygenated state, it tends to form polymers. The polymerization of the defective β-globin (β^S) triggers a process called Sickling, whose hallmark is the generation of abnormal, rigid and hook/sickle-like shaped RBCs that have a shorter than normal life span. Anemia, obstruction of microvessels and periodic pain episodes (crisis) that lead to organ demage are consequences of this condition.

β-thalassemias can be caused by more than 200 β-globin gene (HBB) mutations that affect the synthesis of the β-globin chains, leading to an excess of α-globin chains and an accumulation of free α-subunits. The deficiency of the β-globin chains production can be partial or complete (β^+ and β^0 genotypes, respectively). The phenotypic manifestations vary depending on the severity of the disease, and usually three main types of the disease are considered: β-thalassemia major (severe phenotypic manifestation), β-thalassemia minor (mild phenotypic manifestation) and β-thalassemia intermedia (moderate phenotypic manifestation) (Forget & Bunn, 2013; Cavazzana *et al*. 2017).

1.1.2.1. Current β-hemoglobinopathies Therapies

Therapy available for these hemoglobinopathies remains mostly supportive, consisting in symptomatic care and lifelong regular blood transfusions, which currently are standard treatments for these disorders. However, transfusions need to be combined with chelation therapy to prevent iron overload and organ damage. Bone marrow transplantation involves replacing the disease stem cells in the bone marrow with healthy cells from a legible donor. It can be curative, but the availability of donors is restricted, and the procedures carry significant risks. Allogeneic hematopoietic stem cell transplantation (HSCT) is also an option that relies critically on the availability of matched donors, besides carrying substantial risk of morbidity and mortality (Cavazzana *et al*. 2017; Neumayr *et al*. 2019; Saraf & Rondelli, 2019).

In the last few years, therapeutic approaches aimed at stimulating the production of HbF have attracted great attention. This was due to knowledge that elevated levels of fetal hemoglobin (HbF) ameliorate the clinical severity of β-hemoglobinopathies by reducing globin chain imbalance in β-thalassemias and by inhibiting sickling in SCD (Antoniani *et al.* 2018). Up to now, hydroxyurea has been the standard of care treatment to induce production of HbF.

Hydroxyurea is a small molecule that increases the amount of γ-globin, which in turn promotes the production of HbF. It is used in patients with SCD and in some with β-thalassemia. However, the HbF response to hydroxyurea is variable and not all patients react well to treatment (Steinberg *et al*. 2014).

More recently, genome editing technologies, such as Clustered Regularly Interspaced Short Palindromic Repeat – associated 9 (CRISPR-Cas9), are becoming viewed as a promising therapeutic possibility for β-globin disorders. By targeting HbF repressors, deleting them or reducing their expression, HbF levels could potentially increase. However, transcriptional factors usually regulate multiple target genes, therefore some unwanted side effects can be encountered. The ability to introduce genome modifications can also be used to avoid the action of cis-regulatory ellements, specifically at the γ-globin locus, by mutating their binding sites. This has challenged several laboratories to attempt mimicking hereditary persistence of fetal hemoglobin (HPFH) by introducing known HPFH mutations (Wienert *et al*. 2018).

1.1.3. Genetic Modifiers of Fetal Hemoglobin (HbF)

Fetal hemoglobin (HbF,$\alpha_2\gamma_2$) is the predominant type of hemoglobin during fetal development and constitutes about 60-80% of total hemoglobin at birth. In the postnatal period, HbF rapidly decreases as it is replaced by adult forms of hemoglobin. At around 6-12 months of age, it represents typically <1% of the total hemoglobin in healthy individuals. However, adult inter-individual levels of HbF vary considerably, which is largely determined by genetic variations (Mandal & Kartthik, 2019).

In patients with β-hemoglobinopathies, there is also significant variability in HbF levels, but they remain typically higher when compared to the general population. Part of this variability results from pathological mechanisms associated with the disease. However, genetic factors act on any individual, regardless of whether or not they have some type of pathology, contributing to the observed variability of HbF levels.

High levels of fetal hemoglobin in adulthood are considered to have little physiological and clinical significance in healthy individuals. However, β-hemoglobinopathies patients with higher levels of HbF demonstrate less severe clinical manifestations of the disease (Menzel & Thein, 2018). In addition, clinical observations in infants with SCD and β-thalassemia showed they were asymptomatic until the expression of HbF decreases in the months after birth (Sankaran & Orkin, 2013).

Therefore, understanding the mechanisms of the β-globin gene cluster regulation, particularly how the γ-globin gene is silenced, is of utmost importance for the prognosis and personalized pharmacological treatment of the disease, as well as for the development of new therapeutic strategies that stimulate the production of this beneficial form of hemoglobin. As already emphasized, HbF is now being seen as a highly promising therapeutic target for the treatment of β-hemoglobinopathies.

The hereditary persistence of fetal hemoglobin (HPFH) is an inherited condition characterized by elevated production of γ-globin in adulthood, and consequently, by high HbF levels (~1-5%). Usually, people with HPFH are asymptomatic and appear to be healthy, showing that high levels of HbF are not problematic (Martyn *et al.* 2018). The “Deletional HPFH” is due to large deletions between the γ-globin and the β-globin genes in the β-globin gene cluster, many of them including the loss of the HBB or HBD genes. These deletions might eliminate HbF inhibitory sequences. The “Non-deletional HPFH” is caused by point mutations, or small deletions, in the γ-globin genes promoters. Such

variants are supposed to disrupt binding sites for γ-globin silencers (Cavazzana *et al*. 2017; Weinert *et al*. 2018).

However, these mendelian forms of HPFH are rare and do not explain common HbF variability in normal adults and in patients with β-hemoglobinopathies.

Recent genome-wide association studies (GWAS), conducted in both healthy individuals and patients of European, African or Asian descent affected with β-hemoglobinopathies, have led to identify several commonly occurring variants along the genome associated with HbF levels. Each of these single nucleotide polymorphisms (SNPs) has little impact on HbF levels, but because they act together in a polygenic manner, as a whole they amount to a substantial proportion of the overall HbF variability. Their action is identifiable in population studies as typical quantitative trait loci (QTLs) both in the general populations and in β-hemoglobinopathies patients. When these variants are co-inherited with conditions associated with "stress erythropoiesis", like SCD and β-thalassemia, the HbF levels become amplified, with beneficial clinical effects (Pereira *et al*. 2015; Menzel & Thein, 2018).

Among the most well-established genetic factors associated with HbF levels are included SNPs located at the promoter of the HBG2 gene present in the β-globin gene cluster on chromosome 11, namely the variant rs7482144 also known as XmnI, SNPs at the BCL11A gene on chromosome 2 and SNPs at the HBS1L-MYB (HMIP) intergenic region on chromosome 6 (Pereira *et al.* 2015).

As previously mentioned, the LCR loops to the embryonic globin genes in primitive erythroid cells, leading to dominant expression of ε-globin. At the fetal stage and until birth, the LCR loops to the γ-globin genes allowing their upregulation. Finally, In adult erythroid cells, the LCR interacts with the adult δ-globin and β-globin genes promoters to enable their expression. A pentrameric complex (including GATA1, TAL1, E2A,LMO2, and LDB1) is thought to mediate the formation of the loop between the LCR and the "β-like" globin genes promoters (Figure 4) (Cavazzana *et al.* 2017).

Several transcription factors are implicated in the Hb switching process (Figure 5). The presence of specific repressor proteins during development renders the promoters of the embryonic/fetal globin genes inaccessible for activation, shifting the LCR-promoter contacts to the adult globin genes (Philipsen & Hardison, 2018).

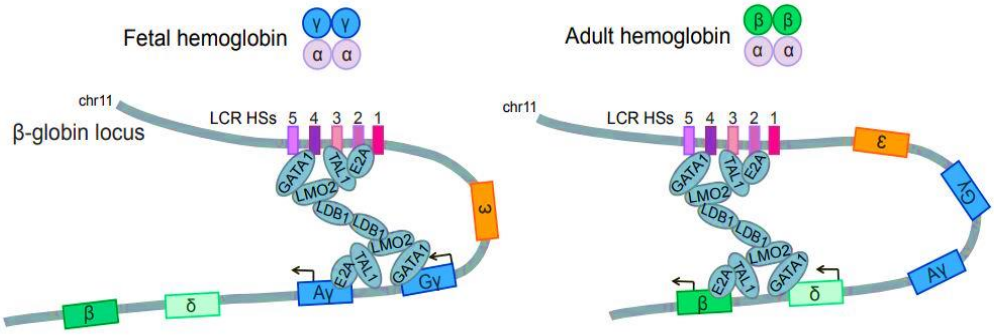


Figure 4 – A pentrameric complex mediates long-range interactions between the LCR and the fetal (blue) and adult (green) globin genes in fetal and adult erythroblasts, respectively (Cavazzana *et al.* 2017).

The B-cell lymphoma 11A protein (BCL11A) is a zinc finger transcription factor with a major role in the silencing of the γ-globin expression. BCL11A seems to bind to a number of sites within the β-globin locus, interacting with the erythroid master regulator GATA1, SOX6, FOG1 and the NuRD chromatin remodeling complex, in order to repress fetal globin gene expression (Cavazzana *et al.* 2017; Habara *et al*. 2017).

The Leukemia/lymphoma-related factor (LRF), encoded by ZBTB7A, is another zinc finger transcription factor recently identified as an important repressor of the γ-globin gene expression in adult erythroid cells. LRF exerts its activity recruiting a unique NuRD repressor complex, independently of BCL11A (Cavazzana *et al.* 2017).

The Kruppel-like factor 1 (KLF1) has been known to participate in the hemoglobin switch. KLF1 is mostly expressed in the adult stages of development and binds with higher affinity to the β-globin promoter than to the γ-globin promoter, preferentially activating the β-globin genes. Adittionaly, KLF1 functions as an indirect repressor of the fetal globin genes, being involved in the activation of the BCL11A expression by binding to its promoter (Siatecka & Bieker, 2011).

In adult erythroid cells, the stage selector protein MYB activates KLF1. Since one of the several major functions of KLF1 is to activate BCL11A and LRF, two transcription factors which act as direct repressors of the fetal globin genes, MYB is thought to indirectly regulates γ-globin transcription. This regulatory circuit results in high-level expression of adult globin genes and efficient repression of fetal genes in adult erythroid cells (Cavazzana *et al.* 2017; Philipsen & Hardison, 2018).

Other nuclear receptors, chromatin modifiers and remodeling factors are also thought to be involved in this complex process (Cavazzana *et al.* 2017).

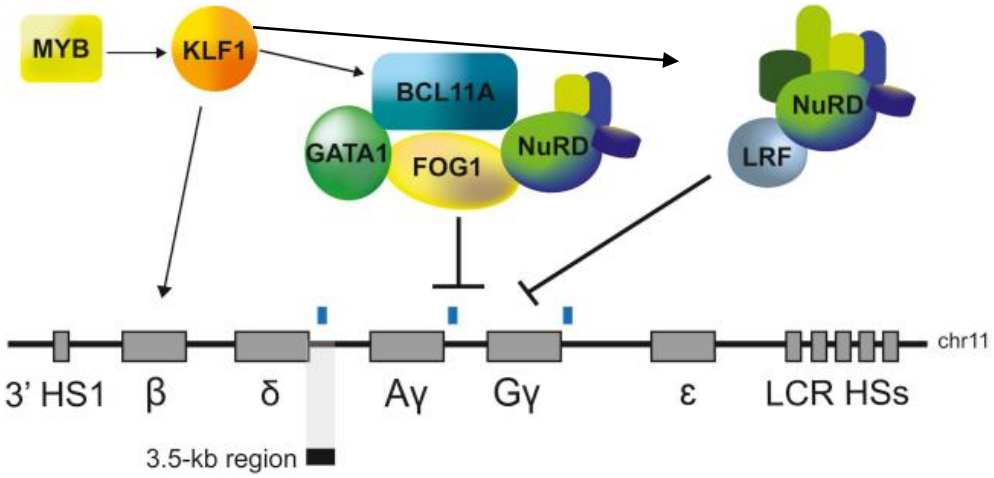


Figure 5 – Binding sites for BCL11A (blue squares) throughout the β-globin gene cluster.BCL11A and LRF are two y-globin genes repressors that function through independent NuRD complexes. Their expression is positively regulated by KLF1, which is upregulated by MYB. Adittionaly, KLF1 favors Hb switching by directly activating the β-globin gene expression (Adapted from Cavazzana *et al*. 2017).

1.1.3.1. The BGLT3 Gene

The Beta Globin Locus Transcript 3 (BGLT3) gene that codes for an erythroid-specific long noncoding RNA (lncRNA) is located in the β-globin cluster, lying in the intergenic region between HBBP1 and HBG1 (Figure 3B).

LncRNAs are emerging as key participants in the regulation of important cellular processes such as gene expression. Since most lncRNAs are highly cell-type specific, hundreds of them being expressed specifically, for example, during erythropoiesis, they may play an important role in the generation of various cell types and in cell-specific functions (Ivaldi *et al.* 2018).

Recently, chromosome conformation capture (3C) enabled to detect looping from BGLT3 sequences to HBg1 and HBG2 in human erythroid cells (Kiefer *et al.* 2011). In addition, it was found that the BGLT3 transcription was consistently associated with γ- globin transcription. Later, the BGLT3 transcript and the γ-globin genes were demonstrated to be co-expressed, during erythroid development and maturation, in an *in vivo* model

(Ivaldi *et al.* 2018). The same study showed that deleting the BGLT3 gene or depleting its transcript decreased γ–globin gene expression, suggesting an activating role of BGLT3. Congruently, the gene locates in a region where naturally occurring deletions are known to cause HPFH.

The region encompassing this gene exhibit enhancer properties and seems to influence γ-globin expression through long-range LCR interactions, including looping from the BGLT3 sequences to the HBG1 and HBG2 genes (Figure 6).

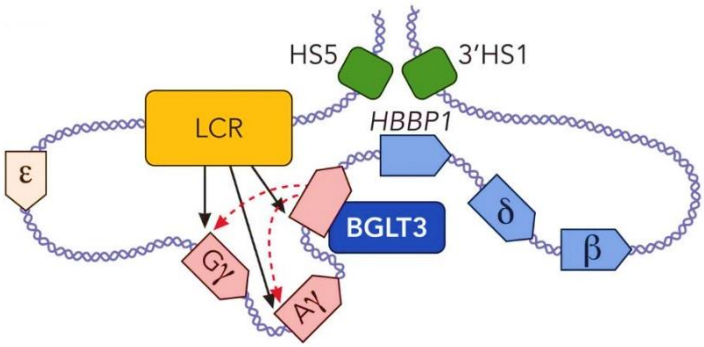


Figure 6 – BGLT3 influences γ-globin expression via its transcript and via an enhancer-like activity, that requires the act of transcription through the BGLT3 gene, contacting the LCR (orange), looping to the γ-globin genes (red) and recruiting RNA Pol II to the promoters of the HBG1 and HBG2 genes (Blobel & Crossley, 2018).

Both the transcription through the BGLT3 gene, as well as the BGLT3 transcript itself were reported as positive regulators of the duplicated γ–globin genes. However, they appear to operate via distinct mechanisms. The BGLT3 locus transcription is crucial for the maintainance of BGLT3 enhancer-like state and for looping between the BGLT3 sequences and the γ-globin genes, while BGLT3 transcript interacts with the mediator of RNA polymerase II transcription (Mediator) complex. Both activities seem to regulate γ-globin transcription in a developmental stage-specific manner, together with the LCR, by serving as a separate means to recruit RNA Pol II to the promoter of the γ-globin genes, contributing to γ-globin gene activation (Ivaldi *et al*. 2018; Blobel & Crossley, 2018).

In a previous study from our group, one SNP (rs7924684) in the BGLT3 gene was found to be strongly associated with high HbF levels in a cohort of β-thalassemia minor patients (Santos, 2019). However, the relationship between genetic variations at this gene and

HbF levels in individuals without hematological disorders is still unknown and so, further studies addressing this issue are necessary.

1.1.3.2. The HBBP1 pseudogene

Besides BGLT3, the intergenic region between the γ–globin and δ–globin genes contains another gene that also produces a noncoding RNA and that is highly expressed in adult erythroid cells: the pseudogene HBBP1.

Up to now, few studies addressed the relationship between genetic variations at HBBP1 and levels of HbF, but the presence of developmental regulatory elements in this region has already been suggested. Two SNPs residing within the second intron of HBBP1, rs10128556 and rs2071348, were previously associated with high HbF levels (Galarneau *et al.* 2010; Nuinoon *et al.* 2010). However, these two SNPs are extremely nearby from each other, being in total or quasi-total linkage-disequilibrium (LD) in most populations, making it difficult to discriminate the influence of each of them in levels of HbF.

Reportedly, the function of HBBP1 is opposed to that of the BGLT3 gene by contributing to fetal globin repression. The β-globin pseudogene appears to control gene expression by influencing the architecture of the β-globin locus. In a recent study, it was found that, at the fetal stage, the HBBP1 region interacts with 3'HS1 and HS5 to separate the δ-globin and β-globin genes from the enhancer LCR, enabling LCR-γ-globin contacts. Contrarily, at the adult stage, the HBBP1 region forms contacts with a region within ε-globin, facilitating LCR interactions with the adult globin genes and decreasing γ-globin transcription (Figure 7). The well-established repressor BCL11A might be implicated in the activity of the HBBP1 region (Huang *et al.* 2017; Blobel & Crossley, 2018).

Contrarly to BGLT3, the HBBP1 gene region, and not the transcript itself or the act of transcription throught the gene, seems to be required for γ-globin silencing in adult erythroid cells.

In short, the BGLT3 and the HBBP1 seem to influence by different still unclear mechanisms the hemoglobin switching process.

In line with this, in the work of Santos (2019), where β-thalassemia carriers were enrolled, the variant rs10128556 in HBBP1 was also found to be significantly associated with HbF levels.

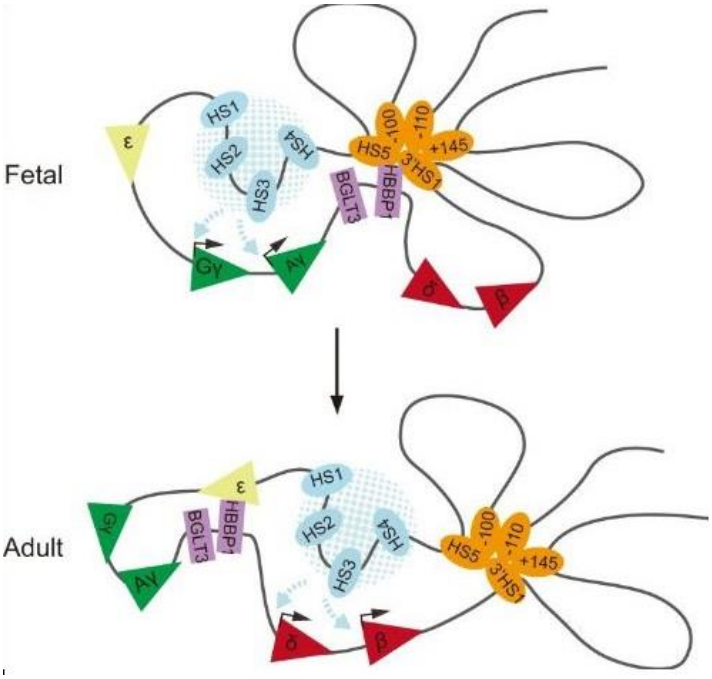


Figure 7 – The HBBP1 region (purple) mediates dynamic chromatin interactions with 3'HS1/HS5 (orange) at the fetal stage and with ε-globin (beige) at the adult stage, influencing chromatin configurations in fetal and adult erythroblasts (Huang *et al.* 2017).

1.2. β-globin Gene Cluster Haplotypes

Mutations in the β-globin gene cluster that cause hemoglobinopathies arise in specific genetic backgrounds within the region. Such backgrounds are defined by the combination of alleles at the polymorphic variations spread throughout the cluster, most of which are single nucleotide polymorphisms (SNPs). SNPs constitute the most abundant class of genetic variation in the genome, occuring, on average, once every a thousand base pairs throughout the genome, with the majority being biallelic. The phenotypic consequences of the SNPs may vary depending on the region they occur (coding or non-coding regions, just for instance).

The specific combinations of alleles at multiple markers defines an haplotype. Each individual has two haplotypes for a specific region of the autosomal genome, representing the maternal and paternal chromosomal contribution. Recombination or mutation create new haplotypes, which can potentially be transmitted to the next generation.

Throughout the years, a growing interest emerged in haplotyping the β–globin gene cluster. The approach revealed to be enlightening to address topics within the scope population and evolutionary genetics, in order to dissect the origin of β-globin gene mutations or to investigate migration patterns and evolutionary relationships between human populations worldwide (Crawford *et al.* 2002). Moreover, as soon as it was demonstrated that the haplotypic backgrounds in which pathogenic mutations are found had clinical implications, haplotypic studies were refoccused to assess wheather haplotypes could assist to predict the clinical severity of β-hemoglobinopathies.

In the first works, the β-globin cluster haplotypes were determined using Restriction Fragment Length Polymorphism (RFLP) based methods. The approach was applied to analyze the sickle cell mutation in the β–globin locus, which was found only in a few haplotypic backgrounds. They become to be referred to as classical HBB*S haplotypes, which totalized five, that were identified based on the presence or absence of restriction recognition sites. Giving their ethnic and/or geographical pattern of distribution, the five haplotypes are mentioned as: Bantu/Central African Republic (BAN/CAR), which is prevalent in South-Central and Eastern Africa; Benin (BEN), dominant in the African Midwest; Senegal (SEN) typical of Atlantic Africa; Cameroon (CAM), usually found within the geographical boundaries of Cameroon and, at a smaller frequency, in the west coast of Africa; and Arab-Indian (AI), predominant in the Arabian Peninsula and India (Table 1) (e.g. Leal *et al.* 2016).

Later, it was discovered that the different haplotypes were associated with varied levels of HbF, and consequently, with different degrees of disease severity. The AI and SEN haplotypes were correlated with high HbF levels, and thus an amelioration of the clinical symptoms of the disease. Contrarily, the Bantu/CAR haplotype was associated with low levels of HbF and a more severe phenotypic manifestation. The BEN and CAM haplotypes were related to intermediate HbF levels and intermediate severity of the disease (Vinson *et al.* 2004).

These findings suggested the existence of cis-acting elements within the β-globin cluster that were modulating HbF expression.

Haplotypes \ Enzymes	*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 patterns of HBB*S associated haplotypes (Couto, 2017; adapted from Bitoungui *et al*. 2015).

Haplotyping of β-thalassemia variants has also become a common practice. After assessing cleavage at seven RFLP sites in β-thalassemia chromossomes of Mediterranean origin patients, nine restriction patterns in the human β-globin gene cluster were found to be the most common (Figure 8).

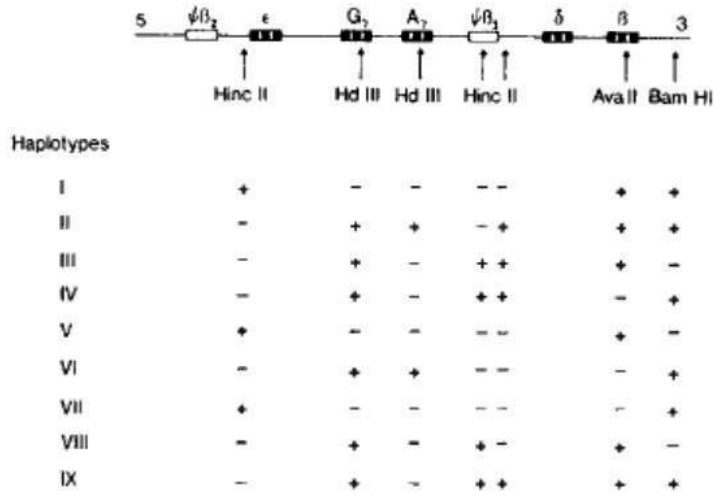


Figure 8 – Representation of the β-globin gene cluster containing the polymorphic restriction sites and the restriction patterns obtained, designated by Haplotypes I-IX (Orkin *et al*.1982).

β-thalassemia haplotypes were also demonstrated to be associated with different levels of HbF. A study showed that the haplotypes IX (in haplotypic homozygotes and heterozygotes) and III (in heterozygotes) were related to high Gγ-globin gene expression and an amelioration of the disease. The haplotypes I and V were associated with low

Gγ-globin expression, low HbF levels and an average clinical course of the disease (Labie *et al.* 1985).

The RFLP-based methods continue to be used to haplotype the β-globin gene cluster. However, because they are time-consuming, a number of alternatives to the conventional methods were developed. Indeed, when later it was realized that most of the restriction sites corresponded to SNPs and such SNPs were identified, direct examination of the SNPs was made possible. Among the strategies devoleped for haplotyping the β-globin gene cluster, is the SNaPshot® Multiplex system implemented by Couto (2017). The system targets the most informative positions defining the SCD haplotypes and contains six SNPs (rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905), which correspond to the restriction sites conventionally examined in RFLP-based approaches. Importantly, the variants rs7482144, located in the promoter region of HBG2, and rs10128556, in the pseudogene HBBP1, are currently two candidate modifiers of HbF levels.

2. Objectives

Despite the accumulated evidence that common polymorphisms at the β-globin gene cluster itself account for the variability in HbF levels, there are still many unknowns on the subject. Which polymorphisms are those? Do they only act in particular haplotypic backgrounds? In which physiological/hematological conditions do they modulate the amount of HbF?

The β-globin cluster contains the pseudogene HBBP1 that was recently shown to be important for γ-globin silencing in adults (Huang *et al.* 2017). It also contains a gene - BGLT3, encoding a developmental stage-specific lncRNA that was demonstrated to positively regulate γ-globin genes (Ivaldi *et al.* 2018). However, there are still very few studies addressing the role of variations at BGLT3 and HBBP1 as genetic modifiers of HbF levels in adulthood. In a recent work from our team (Santos, 2019), analysis of diversity at different SNPs encompassing the entire β-cluster allowed to detect significant associations with levels of HbF in a sample of Portuguese carriers of β-thalassemia.

The purpose of this study is to enlarge the previous analysis to a sample of subjects with similar population ancestry but lacking any hematological disorder. To this end, and similarly to the strategy before used by Santos, we will be taking advantage of the Multiplex SNaPshot® system for haplotyping the β-cluster recently developed by Couto (2017), which contains 6 SNPs, including one in HBBP1 (rs10128556). Using this 6-plex system as the starting tool, the goal is to add it with rs7924684 located in the BGLT3, because according to Santos, who examined the gene through direct sequencing, it was the unique SNP within BGLT3 with major allele frequency >0.01 that in addition was one of those significantly associated with HbF levels in carriers of β-thalassemia.

Analyzing now a sample of subjects not affected by hematological diseases, the major aim of the work is to evaluate whether the role of common polymorphisms at the β-globin cluster as modulators of HbF is dependent on possible physiological stress responses experienced by β-thalassemia minor patients. Ultimatedly, we hope to provide novel insights on the β-globin cluster that might be useful to the development of therapeutic approaches that induce HbF in order to ameliorate the severity of β-hemoglobinopathies.

3. Materials and Methods

3.1. Sampling

In this study, 51 DNA samples from Portuguese unrelated subjects (32 females and 19 males) were analyzed. All provided informed consent to participate in the study.

The individuals here examined had already been enrolled in a study adressing the influence of polymorphic variations at SNPs located in the loci BCL11A (2p16), HBS1L-MYB (6q23), and in the HBG2 promoter (11p15.5) in levels of HbF (Pereira *et al.* 2015).

According to the available hematological data, HbF levels ranged from 0.1% to 9.1%, meaning that some subjects had an increased HbF level, overcoming thus the normal range. HbF and HbA2 levels were determined by high performance liquid chromatography using Variant 2 (Pereira *et al.* 2015; Bio-Rad, CA, USA).

3.2. 7-Plex SNaPshot® System

For haplotyping the β cluster, it was applied the Multiplex SNaPshot® system implemented by Couto (2017) for HBB*S haplotyping. Couto identified the fragments typically amplified and submitted to restriction enzymes in the conventional HBB*S haplotyping methodology - Restriction Fragment Length Polymorphism (RFLP) based methods – and was able to determine the precise location of the most informative polymorphisms defining the haplotypes, which led to the selection of the six target SNPs (rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905) then implemented in the system.

In this study, one SNP located in the BGLT3 gene (rs7924684) that showed to be associated with high HbF levels in β-thalassemia minor patients (Santos, 2019) was added to the system, developing thus a 7-plex SNaPshot strategy (Figure 9).

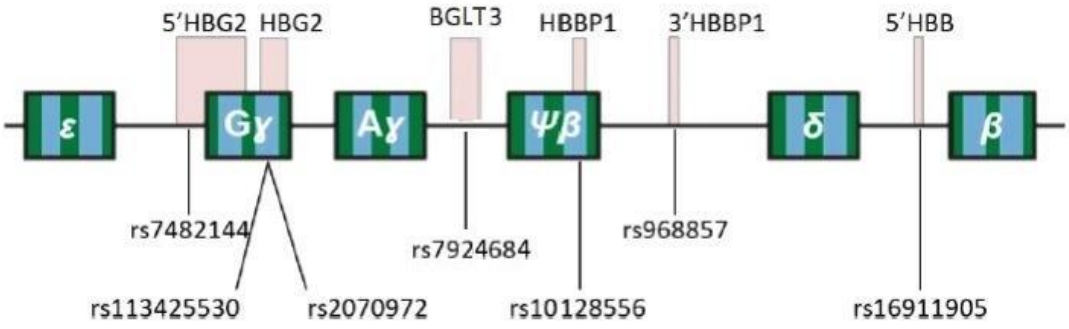


Figure 9. Representation of the β-globin cluster, showing the genomic regions to be amplified (pink squares) that carry the target SNPs (adapted from Couto, 2017).

3.2.1. Multiplex PCR

The desired fragments were amplified by Multiplex PCR using five pairs of primers designed by Couto (5'HBG2_FW/RV; HBG2_FW/RV; HBBP1_FW/RV; 3'HBBP1_FW/RV; 5'HBB_FW/RV) and one pair of primers designed by Santos (BGLT3_FW/RV) (Table 2).

Primer	Sequence (5'>3')	Fragment Lenght (bp)
5'HBG2_FW	ACAAGAAGGTGAAAAACGG	772
5'HBG2_RV	CTTTATGGCATCTCCCAAG	
BGLT3_FW	GCCACAAACAAGAAAGAATC	538
BGLT3_RV	CCCCATGTGTATTTCAAGGA	
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 2. Primers used for amplification of the target regions of the β-globin cluster and expected fragment sizes, which were intentionaly designed to differ in size from each other to possibilitate the implementation of the Multiplex PCR.

Multiplex PCR was performed in a final volume of 5 µL per reaction, 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), 1.5 µL of distilled water and 0.5 µL of DNA. Negative controls with 0,5 µL of distilled water instead of DNA were used. In Table 3 are specified the Multiplex PCR conditions implemented. Amplification was carried out in a T100TM Thermal Cycler (Bio-Rad).

Step	Temperature (ºC)	Time	Cycles
Initial Denaturation	95	15 min	1
Denaturation	94	0:30 min	
Annealing	60	1:30 min	35
Extension	72	1 min	
Final Extension	72	10 min	1
Hold	12		∞

Table 3. Multiplex PCR protocol used for amplification of the target regions of the β-globin cluster.

To control the success of the amplification reaction, the PCR products were submitted to 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.

In gel wells, it was loaded 1.2 µL of PCR amplification products and 1.2µL of 0'GeneRuler 100 bp Plus DNA Ladder (Thermo ScientificTM). For the electrophoretic run, it was applied a constant voltage of 180V during aproximatedly 40 minutes.

The DNA fragments were visualized by the Silver Staining method, which consisted in submitting the gel to the following steps: 10 minutes in Ethanol (10%); 5 minutes in Nitric Acid (1%) with gentle agitation, followed by 2 washings with destilled water; 20 minutes in Silver Nitrate (0.2%) with gentle agitation, in the dark, followed by 2 washings with distilled water; DNA visualization solution composed of 6g of Sodium Carbonate (0.28M), 1mL of Formaldehyde (4%) and 200mL of destilled water; 2 minutes in Acetic Acid for the termination of the reaction and a final washing with distilled water.

3.2.2. SNaPshot®

For the SNaPshot strategy, six Single Base Extension (SBE) primers were designed by Couto (2017) to hybridize with the sequences immediately adjacent to the target SNPs. Another SBE primer was here designed to target the SNP present in the BGLT3 gene that was also implemented in the strategy. Non-annealing tails with different sizes were placed at the 5' end of each SBE primer as presented in Table 4, to allow the discrimination by size and the analysis of multiple SBE products generated from SNaPshot.

rs	Polymorphism	Detected Allele	SBE Sequence (5'>3')	bp
7482144	G>A	G/A	GGTGGAGTTTAGCCAGG	18
113425530	C>A	G/T	GTCGTGAAAGTCTGACAATTGATTCTGGGTGGAA	35
2070972	C>A	C/A	GACTAAACTAGGTGCCACGTCGTGAAAGTCTGACAAC CTCCAGATAACTACACACC	57
7924684	C>T	G/A	TAAACTAGGTGCCACGTCGTGAAAGTCTGACAATGAT GGAAAGATGGGGCA	52
968857	T>C	T/C	CAATGCATGACACATGCTTG	21
16911905	G>C	G/C	TGCCACGTCGTGAAAGTCTGACAACGTTTTAAAATCA TTTCCTT	45
10128556	C>T	G/A	GTCTGACAATGTTGGGGTAGTGAGTTG	28

Table 4. Single Base Extension (SBE) primers used for the genotyping on the target polymorphisms.

An enzymatic cleanup of the PCR products was performed to eliminate the remaining PCR primers and unincorporated dNTPs. Viewing that it was added 0.5 µL of ExoSAP-IT™ (Applied Biosystems™) to 1 µL of amplification products and applied the temperature conditions presented in Table 6 – Initial Purification.

To each purified product, it was added 1µL of SNaPshot™ Multiplex Kit (Applied Biosystems™), 1 µL of SBE Primer Mix and 1.5 µL of destilled water. The SBE Primer Mix was prepared in order to obtain the concentrations in the final solution displayed in Table 5. The SNaPshot® reactions took place under the conditions described in Table 6 - SNaPshot®. The reaction promotes the annealing of the SBE primers to the target sequences followed by the incorporation of a labeled dideoxynucleotide (ddNTP) complementary to each SNP by AmpliTaq® DNA Polymerase.

SBE	Concentration in final solution (5µL)
rs7482144	0.8 µM
rs113425530	1.9 µM
rs2070972	0.3 µM
rs7924684	1.2 µM
rs10128556	0.3 µM
rs968857	0.8 µM
rs16911905	0.6 µM

Table 5. Single Base Extension (SBE) primers concentrations used in the SBE Primer Mix.

A final purification step to clean the SBE products of unused dye-linked ddNTPs (Table 6 – Final Purification) was performed by adding 1 µL of FastAP (Thermo Scientific™) to the final product.

All the reactions were conducted in a T100TM Thermal Cycler (Bio-Rad).

Initial Purification	
Temperature (ºC)	**Time**
37	15 min
85	15 min

SNaPshot®		
Temperature (ºC)	**Time**	**Cycles**
96	10 sec	25
50	5 sec	
60	30 sec	

Final Purification	
Temperature (ºC)	**Time**
37	1 h
85	15 min

Table 6. Protocols used for the SNaPshot® strategy.

Lastly, 9.5 µL of a mixture of Hi-Di™ Formamide (Applied Biosystems™) and GeneScanTM – 120 LIZ™ Size Standard (Applied Biosystems™) were added to 0.5 µL of the SNaPshot® purified products. After capillary electrophoresis, in a 3500 Genetic Analyzer (Applied Biosystems™), for separation of the prducts, the results were analyzed with the GeneMapper® Software 5.

3.2.3. Haplotype Inference

The polymorphic variations defining the conventional HBB*S haplotypes, Bantu/CAR, Benin, Senegal, Arab-Indian and Cameroon are presented in Table 7.

Three distinct haplotypes can be associated to the RFLP-based Benin configuration since rs113425530 and rs2070972 are located in the same restriction enzyme recognition site. This way, when RFLP analysis is performed, the presence of nucleotide T in the first SNP (rs113425530) or C in the second (rs2070972) promotes enzyme recognition (cut/+), while non-recognition occur in the presence of G an A in rs113425530 and rs2070972 , respectively (absence of cut/-).

In the sample here characterized, haplotype inference was performed using the Expectation Maximization/Excoffier-Laval-Balding (EM/ELB) algorithm implemented in the software Arlequin v.3.5.

	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
Cameroon	G	G	A	G	T	G

Table 7. Alleles of 6 Single Nucleotide Polymorphisms defining the main sickle cell haplotypes.

3.3. Statistical Analysis

Allele frequencies in the different SNPs were estimated by direct counting with Arlequin v.3.5 (Excoffier *et al.* 2010). The same software was used to test the Hardy-Weinberg equilibrium and to conduct haplotype inference through the Expectation Maximization/Excoffier-Laval-Balding (EM/ELB) algorithm there implemented.

Analysis and visualization of Linkage Disequilibrium (LD) was performed with the Haploview software (Barrett *et al.* 2005).

The software PLINK (available at https://zzz.bwh.harvard.edu/plink/download.shtml) (Purcell *et al*. 2017) was used to evaluate the association between SNPs and levels of

HbF. Two different statistical approaches were undertaken: I) linear regression under an additive genetic model was executed when only subjects with normal HbF levels were considered; the test was done after logarithmic transformation of HbF values; II) a case-control association test was carried on assuming subjects with HPFH vs subjects with normal HbF levels; 2% HbF was used as cutoff-point, crude p-values and adjusted p-values for age and sex as covariates were obtained.

Graphical analyses, normality of the data calculated by the Kolmogorov-Smirnov test and comparisons of HbF levels between genotypes by using one-way-ANOVA were performed with the IBM® SPSS® Statistics software.

4. Results and Discussion

Demographic and hematological data of the 51 portuguese healthy individuals (19 males and 32 females) analyzed in this work are shown in Table 8. Among subjects with HbF levels >5% the presence of deletional mutations in β-genes had been previously excluded through molecular testing.

Characteristics of the sampled individuals	
Age (mean ± SD, in years)	37 (± 16,9)
Age (range, in years)	2 -- 75
Males (n)	19
Females (n)	32
HbF (mean ± SD; %)	1,65 (± 1,33)
HbF (range; %)	0,1 -- 9,1
HbA2 (mean ± SD; %)	2,7 (± 0,23)
HbA2 (range; %)	2 -- 3,4

Table 8. Demographic characteristics and hematological parameters of the sampled subjects.

4.1. Development of a 7-Plex SNaPshot® System

We have succeed to incorporate rs7924684 from BGLT3 in the SNaPshot® System previously developed by Couto viewing to haplotype the β-globin cluster. In this way, the updated 7-plex SNaPshot® System target now the following SNPs: rs7482144, rs113425530, rs2070972, rs7924684, rs10128556, rs968857 and rs16911905.

In Figure 10 are shown the stained PCR products obtained through multiplex amplification after having been submitted to a polyacrilamide electrophoretic run. The gel illustrated contains the results for 20 of the 51 studied individuals. In turn, in Figure 11 are presented examples of electropherograms plotting the SNaPshot® reactions obtained in two subjects. Complete genotypic data for all individuals is available in Table S1.

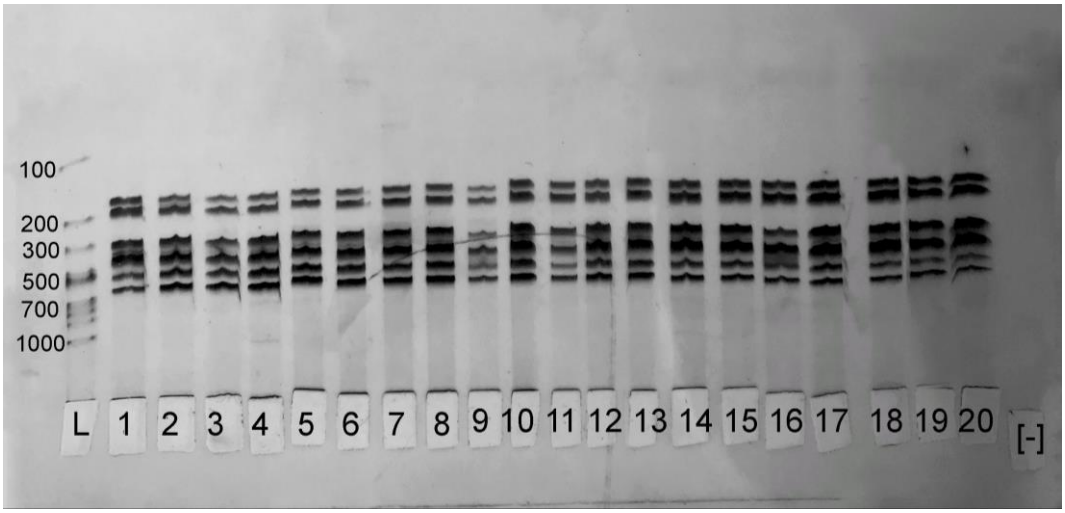


Figure 10. Polyacrilamide gel showing the results of the Multiplex PCR targeting the 7 SNPs. **L-** Ladder; 1- Sample 2493; **2-** Sample 2750: **3-** Sample 2751; **4-** Sample 2779; **5-** Sample 3332; **6-** Sample 3383; **7-** Sample 3939; **8-** Sample 4608 ; **9-** Sample 4812 ; 10- Sample 6308; 11- Sample 6326 **12-** Sample 6903; **13-** Sample 6939; **14-** Sample 7000; **15-** Sample 7033; 16- Sample 7046; **17-** Sample 7503;**18-** Sample 8033; 1**9-** Sample 8519; **20-** Sample 8533; **[-] -** Negative control.

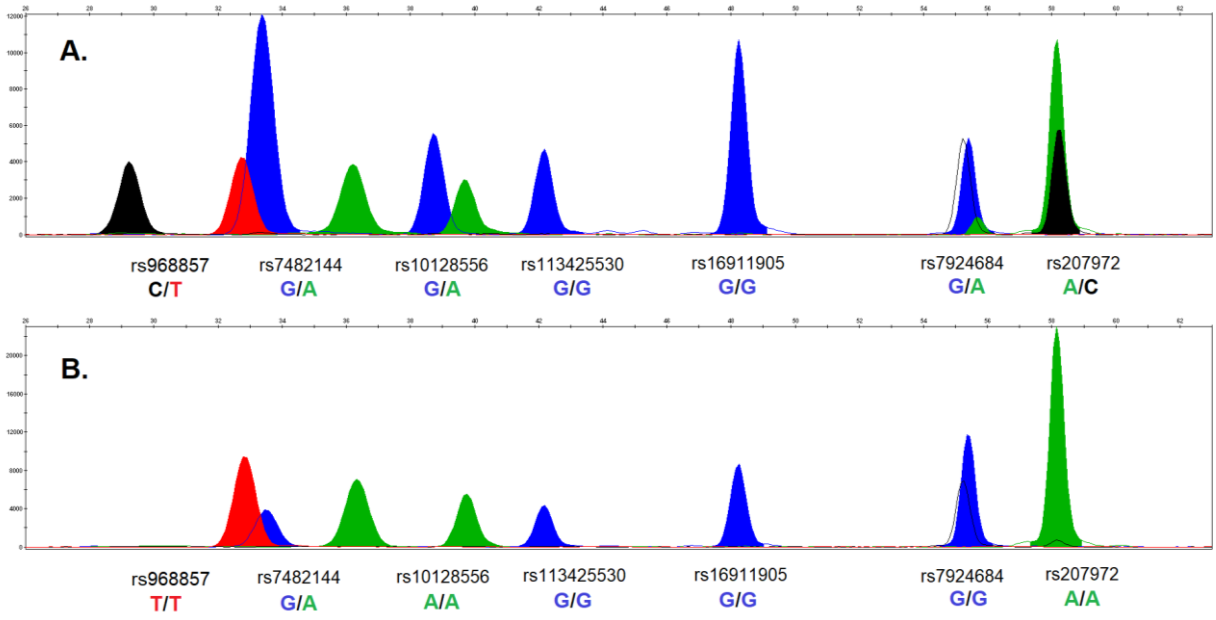


Figure 11. Examples of SNaPshot reaction results. **A:** Sample 3383; **B:** sample 4608.

4.2. Population Data

Using the Arlequin software, allelic frequencies were calculated for each SNP and all values obtained fell in the typical range of values observed in European populations

according to the information available in the Ensembl platform. Hardy-Weinberg Equilibrium (HWE) tests were also performed for all polymorphisms to test the deviations of the observed genotype frequencies from those expected. Genotype distributions in the 7 SNPs were all in conformity with HWE (P-HWE>0,05). The minor allele frequencies (MAF) and P-values for HWE tests (P-HWE) are presented in table 9, for all SNPs under study.

Gene	SNP	N	Alleles (1:2)	MAF	P-HWE	S.D
HBG2	rs7482144	51	A:G	0,333	0,126	0.00033
HBG2	rs113425530	51	T:G	0,118	0,518	0.00051
HBG2	rs2070972	51	C:A	0,431	0,779	0.00041
BGLT3	rs7924684	51	A:G	0,353	0,545	0.00051
HBBP1	rs10128556	51	A:G	0,373	1	0.00000
Intergenic region between HBBP1-HBD	rs968857	51	C:T	0,422	0,149	0.00038
HBB	rs16911905	51	C:G	0,069	0,198	0.00038

Table 9. SNPs targeted in the SNaPshot® Multiplex System, gene/region where they locate and number of samples analyzed (N). Detected alleles (1- minor; 2- major), Minor allele frequencies (MAF) and P-value for Hardy-Weinberg Equilibrium (P-HWE) for each polymorphism with standard deviation (S.D).

Based on the genotypic data, haplotypes were inferred restricting the analysis to the 6 SNPs that define the conventional HBB*S haplotypes (rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905), which lead to detect 13 different haplotypes in the sample (Table 10).

Three of those haplotypes corresponded to conventional HBB*S haplotypes (Table 7). The haplotypes 2, 3 and 5 matched the Arab-Indian (AI), Senegal (SEN) and Cameroon (CAM) haplotypes, respectively. In total, they summed up 44 occurrences, representing 43.1% of the entire set of haplotypes. The SEN haplotype, which has been correlated with high levels of HbF, was the second most common haplotype in the sample.

Haplotype ID	Haplotype	Frequency	Observations
1	GGCGCG	36	
2	AGAATC	4	Arab-Indian (AI)
3	AGAATG	28	Senegal (SEN)
4	GGAGCG	3	
5	GGAGTG	12	Cameroon (CAM)
6	GTCGTG	5	
7	GTAGTG	5	
8	AGAACG	1	
9	AGCATG	1	
10	GGCGCC	2	
11	GTAATG	1	
12	GGAATG	3	
13	GTAGCC	1	

Table 10. Haplotypes in the 51 studied samples deduced by Arlequin software, frequency and corresponding conventional HBB*S haplotypes. Order of SNPs: rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905.

4.3. Linkage Disequilibrium Analysis

The seven SNPs analyzed in this study are located throughout a great extension of the entire β-globin cluster. It has long been known that the cluster presents two distinct regions with strong LD: one containing HBB (LD region 1) and the other encompassing the stretch from HBD to the LCR (LD region 2). The regions are separated by a recombination hotspot (e.g. Moleirinho *et al.* 2013 and references therein). The global pattern of LD in the β-globin cluster is illustrated in Figure 12, where are also presented the relative positions of the seven SNPs here examined.

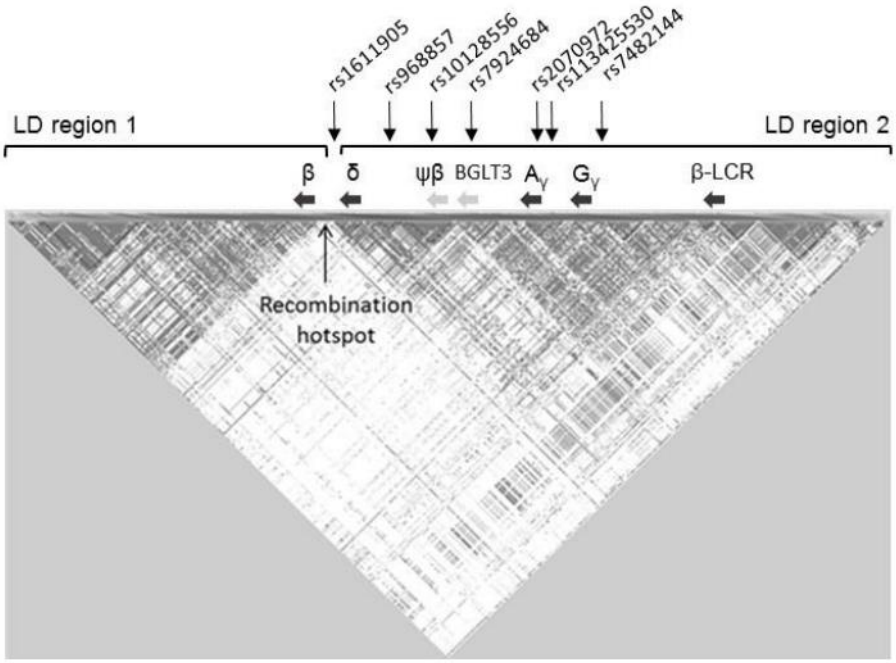


Figure 12. LD plot of the the β-globin cluster for the data from 1000 Genomes Project for YRI population (Yoruba from Ibadan in Nigeria) and location of 7 SNPs in the cluster (Santos, 2019; adapted from Moleirinho *et al*. 2013).

Haploview pairwise analysis of linkage disequilibrium (LD) between the seven SNPs in the population here studied is presented in Figure 13. Haplotype blocks were defined with the confidence interval method of Gabriel *et al.* (2002) and the r^2 measure was used. According to the selection criteria, one LD block involving 3 SNPs (rs968857, rs10128556 and rs7924684) stood out. Two of these SNPs, the rs968857 (located in the intergenic region between HBBP1 and HBD) and the rs7924684 (located in the BGLT3 gene), were in strong LD (r^2=0.74). In addition, one SNP from this block,rs10128556, located in the HBBP1 pseudogene, also showed to be in very strong LD (r^2=0.84) with another SNP not belonging to the block, rs7482144, that is located in the promoter region of HBG2.

The LD pattern here observed was very similar to the one detected by Santos (2019) in the portuguese β-thalassemia carriers then analysed. Considering the relative positions of the SNPs in the β-globin cluster, the unusual LD pattern suggests that besides the physical distance between SNPs, interactions between sequences where they reside must also account for the level of non-random association between alleles at different SNPs.

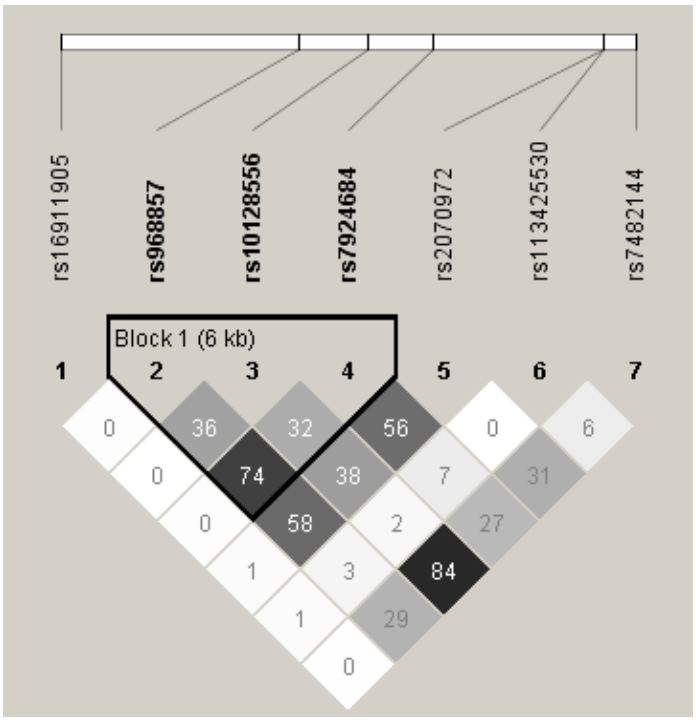


Figure 13. LD plot indicating r^2 values between the 7 SNPs, which are displayed, in percentage, inside de boxes. The schematic representation of LD Blocks was constructed using the Haploview 4.1 software. The r^2 colour scheme was used: white squares, $r^2 = 0$; Squares with different shades of gray, $0 < r^2 < 1$; Black squares, $r^2 = 1$).

4.4. Association Analysis with HbF Levels

In order to assess the relationship between variation at the 7 SNPs of β-cluster and HbF levels, different analyses were conducted considering distinct criteria to group subjects within the total of the 51 Portuguese examined.

The individuals were first divided into two groups according to levels of HbF: the first group (n=36) included the participants with normal HbF levels (≤1.6%) and the second group (n=15) the subjects with higher HbF levels, hereafter designated, for simplicity, as having HPFH (>1.6%). A case-control study was then performed under multiple genetic models.

When an additive model was used, the case-control study showed nominal significant association between rs113425530 SNP and HbF levels (OR=4.08; 95% CI, 1.79-14.1; P=0.038), while not for the remaining SNPs (Table 11). Using sex and age as covariates

did not altered the results. The second highest odd ratio reached 2.18 and was exhibited by rs2070972. That OR was associated to a low P-value, but yet not reaching statistical significance (OR=2.18; 95% CI, 0.92-5.18; P=0.084). Even so, diferent models were next used to further investigate the correlation between these two SNPs and HbF levels.

Chr: position	Gene	SNP	N	MAF (HbF ≤ 1.6%) (n=36)	MAF (HbF > 1.6%) (n=15)	OR (CI 95%)	P	P*
11:5228060	HBB	rs16911905	51	0.083	0.033	0.38 (0.04-3.29)	0.670	0.670
11:5239228	Intergenic HBBP1-HBD	rs968857	51	0.403	0.467	1.30 (0.55-3.06)	0.661	0.661
11:5242453	HBBP1	rs10128556	51	0.403	0.3	0.64 (0.26-1.58)	0.375	0.375
11:5245498	BGLT3	rs7924684	51	0.347	0.367	1.09 (0.45-2.64)	1	1
11:5253487	HBG2	rs2070972	51	0.375	0.567	2.18 (0.92-5.18)	0.084	0.084
11:5253490	HBG2	**rs113425530**	51	0.069	0.233	4.08 (1.79-14.1)	**0.038**	**0.038**
11:5254939	HBG2	rs7482144	51	0.361	0.267	0.64 (0.25-1.65)	0.490	0.490

Table 11. HbF association results for seven SNPs of β-cluster in individuals of Portuguese origin without hematological disorders. The association was tested performing a case-control study under an additive genetic model. Significant association P-values (P<0.05) are in bold.

The odds ratio (OR) shown regard the minor allele. Confidence Interval (CI) 95% and P-values for allelic association were obtained with Fisher's test. P: P-value unadjusted; P*: P-value using age and sex as covariates;

The dominant genetic model (genotypes homozygous + heterozygous for the 'incresing' allele vs homozygous for the other allele) yielded an even higher OR and lower P-value for the rs113425530 (OR=7; 95% CI, 1.64-29.9; P=0.009). Again, no statistical significant OR was produced for the rs2070972 polymorphism (Table 12). However, the case-control study under a recessive model (genotypes homozygous for the 'incresing' allele vs homozygous + heterozygous for the other allele) showed nominal statistical significance for the rs2070972 SNP (OR=5.33; 95% CI, 1.23-23.1; P=0.025) but not for rs113425530 (Table 13). Using sex and age as covariates did not significantly altered the results.

Chr: position	Gene	SNP	N	MAF (HbF ≤ 1.6%) (n=36)	MAF (HbF > 1.6%) (n=15)	OR (CI 95%)	P	P*
11:5228060	HBB	rs16911905	51	0.083	0.033	0.44 (0.05-4.15)	0.476	0.396
11:5239228	Intergenic HBBP1-HBD	rs968857	51	0.403	0.467	1.06 (0.27-4.11)	0.935	0.897
11:5242453	HBBP1	rs10128556	51	0.403	0.3	0.65 (0.19-2.19)	0.483	0.519
11:5245498	BGLT3	rs7924684	51	0.347	0.367	0.65 (0.19-2.19)	0.483	0.531
11:5253487	HBG2	rs2070972	51	0.375	0.567	1.55 (0.41-5.88)	0.516	0.379
11:5253490	HBG2	**rs113425530**	51	0.069	0.233	7 (1.64-29.9)	**0.009**	**0.028**
11:5254939	HBG2	rs7482144	51	0.361	0.267	0.65 (0.19-2.19)	0.483	0.519

Table 12. HbF association results for seven SNPs of β-cluster in individuals of Portuguese origin without hematological disorders. The association was tested performing a case-control study under a dominant genetic model. Significant association P-values (P<0.05) are in bold.

The odds ratio (OR) shown regard the minor allele. Confidence Interval (CI) 95% and P-values for allelic association were obtained with Logistic regression analysis. P: P-value unadjusted; P*: P-value using age and sex as covariates;

Chr: position	Gene	SNP	N	MAF (HbF ≤ 1.6%) (n=36)	MAF (HbF > 1.6%) (n=15)	OR (CI 95%)	P	P*
11:5228060	HBB	rs16911905	51	0.083	0.033	$1.44x10^{-9}$ (0-inf)	0.999	0.999
11:5239228	Intergenic HBBP1-HBD	rs968857	51	0.403	0.467	2.75 (0.49-15.53)	0.252	0.193
11:5242453	HBBP1	rs10128556	51	0.403	0.3	0.36 (0.04-3.26)	0.361	0.475
11:5245498	BGLT3	rs7924684	51	0.347	0.367	4.25 (0.63-28.6)	0.137	0.071
11:5253487	HBG2	**rs2070972**	51	0.375	0.567	5.33 (1.23-23.1)	**0.025**	**0.03**
11:5253490	HBG2	rs113425530	51	0.069	0.233	$1.44x10^{-9}$ (0-inf)	0.999	0.999
11:5254939	HBG2	rs7482144	51	0.361	0.267	$1.36x10^{-9}$ (0-inf)	0.999	0.999

Table 13. HbF association results for seven SNPs of β-cluster in individuals of Portuguese origin without hematological disorders. The association was tested performing a case-control study under a recessive genetic model. Significant association P-values (P<0.05) are in bold.

The odds ratio (OR) shown regard the minor allele. Confidence Interval (CI) 95% and P-values for allelic association were obtained with Logistic regression analysis. P: P-value unadjusted; P*: P-value using age and sex as covariates;

Both these SNPs are located in the same intronic region of the HBG2 gene, and even though being extremely nearby from each other, LD analysis showed that they were not in LD. Although being two of the SNPs defining the conventional HBB*S haplotypes, residing in the recognition site of HindIII, little evidence exists regarding their influence in HbF levels. In a study involving patients from Thailand with HbE/β0-thalassemia, the rs2070972 was associated with HbF levels (Sherva *et al.* 2010). However, to the best of

our knowledge none of the SNPs was reported to influence HbF levels in normal subjects.

Several binding sites for transcription factors involved in fetal hemoglobin regulation were found in noncoding regions of the γ^G-globin gene (Carrocini *et al.* 2015). Speculating on the role of these polymorphisms, the rs113425530 and rs2070972 SNPs may be located in some of such noncoding regulatory regions, being important for the γ-to-β globin switching process. However, it is not possible for now to discard neither that the associations detected with HbF levels were population' specific nor that they were correlated with another not yet identified SNP, possibly in LD with rs113425530 and/or rs2070972, that actually has the really impact on the regulation of γ-globin genes.

It should be noted that applying the Bonferroni correction for multiple testing in all tests performed in the case-control study, assuming thus a $P < 7.1 \times 10^{-3}$, no OR at any SNP was found to surpass the confidence level, meaning that the here detected associations trends with HbF levels still need to be statistically strengthened.

However, we can not ignore the findings since the absence or associations with stronger statistical suport may be due to the limited number of individuals integrating the group with HPFH (n=15) when compared to the number of subjects in the control group with normal HbF levels (n=36).

Next, we centered attention in the subgroup (n=39) of subjects with normal HbF levels (<2%). Values of HbF were converted via logarithmic transformation in order to normalize the quantitative trait distribution and then linear regression analysis, under an additive model, was performed.

In table 14 are presented allele frequencies for each SNP, P-values of the HWE tests (P-HWE) as well as the results of the association analysis performed. The unique SNP that deserved attention was rs968857 that, even so, only revealed a marginal association with HbF levels, once the P-value of the corresponding association test almost reached the non-significant level (β= 0.193; P=0.047). Using sex and age as covariates did not significantly altered the results.

Chr: position	Gene	SNP	N	Alelles (1:2)	MAF	P-HWE	β (SE)	P	P*
11:5228060	HBB	rs16911905	39	C:G	0.077	0.187	0.014 (0.135)	0.919	0.828
11:5239228	Intergenic HBBP1-HBD	**rs968857**	39	C:T	0.423	0.099	0.193 (0.094)	**0.047**	**0.043**
11:5242453	HBBP1	rs10128556	39	A:G	0.397	1	-0.140 (0.081)	0.091	0.060
11:5245498	BGLT3	rs7924684	39	A:G	0.359	0.294	0.153 (0.093)	0.109	0.112
11:5253487	HBG2	rs2070972	39	C:A	0.397	0.523	0.133 (0.086)	0.131	0.152
11:5253490	HBG2	rs113425530	39	T:G	0.064	0.127	-0.154 (0.140)	0.280	0.375
11:5254939	HBG2	rs7482144	39	A:G	0.359	0.294	-0.140 (0.093)	0.142	0.109

Table 14. HbF association results for the seven SNPs in individuals with normal hematological parameters of Portuguese origin (HbF<2%).

The table includes the effect sizes of the minor allele (regression coefficient beta, β), standard error (SE) and P-values for the log transformed HbF levels using a linear regression model. P: P-value unadjusted; P*: P-value using age and sex as covariates;

Using the same model, Santos (2019) obtained very different results when evaluating the association between the same set of SNPs and HbF levels in a sample (n=71) of Portuguese β-thalassemia minor patients, among whom HbF ranged from 0.2% to 8.6%. In that study, four SNPs (rs7482144, rs7924684, rs10128556 and rs968857) were found to be strongly associated with HbF levels ($P < 4.1x10^{-4}$). Interestingly, the strongest signal was displayed by rs968857 (β=0.7; $P=7.091x10^{-6}$), which is the one here identifyied as possibly accounting to the normal range of HbF .

The rs968857 SNP is located in a DNA segment between HBBP1 and HBD, a region also encompassed by the 7.2 Kb Corfu deletion, a thalassemia mutation that in homozygous patients was associated with overwhelming expression of HbF. This is a region that contains BCL11A binding sites. In a recent study envolving sequencing of the region encompassing the Corfu deletion, rs968857 was also shown to be associated with HbF levels in patients with SCD (Akinsheye *et al.* 2011). In that work, in silico analysis revealed that the major allele (T) of this SNP eliminated the AP-1 and NF-E2 binding sites, which were both present in the minor allele (C). These are transcription factors involved in the hemoglobin switching process. Consistently, the Bantu/CAR haplotype, which is associated with low levels of HbF, is the only conventional HBB*S haplotype characterized by the minor allele at this SNP. However, any functional significance of this SNP remains to be explored (Akinsheye *et al.* 2011).

In this association analysis involving only individuals with normal levels of HbF (<2%), the rs113425530 and rs2070972 that in the case-control study firstly presented (where individuals with HPFH were also considered) were significantly correlated with HbF, did not showed similar correlation. To clarify this inconsistency, enlargement of the low number of subjects up to now analyzed can be crucial.

Contrarily to the results obtained here, the rs7482144 and the rs10128556 SNPs were previously found to account for variability in HbF levels. In fact, the T allele of the rs7482144 polymorphism (XmnI), located at the promoter of the HBG2 gene, has been associated not only with higher HbF levels among subjects under erythropoietic stress, namely in β-thalassemia and SCD patients (Pereira *et al*. 2015), but also with a modest increase of HbF in normal individuals in a number of association studies, including for instance those conducted with normal individuals from Algerian and Tunisian origin (Zertal-Zidani *et al.* 2002; Jouini *et al.* 2011).

Still, our results concerning rs7482144 were not unexpected since essentially the same subjects here analyzed were used by Pereira *et al* (2015) in a study where the association with this SNP and HbF levels was already investigated without having found out any statistically significant correlation. In addition, such non-association is in accordance with previous reports indicating minor or absent consequences of variability at rs7482144 on HbF levels in normal individuals of Italian and Indian populations (Guida *et al*. 2006; Dabke *et al.* 2013). These confliting observations, denote that the mechanism whereby this SNP influences HbF is still highly undeciphered.

As for rs10128556, residing within the second intron of the HBBP1 pseudogene, Galarneau *et al.* (2010) found it to be strongly associated with HbF levels, even more than the XmnI-HBG2 (rs7482144) SNP, among African Americans with SCA. A very recent study was able to demonstrate such correlation in normal individuals from Chinese origin (Hu *et al.* 2020). Since the SNP locates in a region belonging to a non-coding sequence, the authors hypothesized that its effect on the HbF expression could result from changing the binding sites of the transcription factors that are related to the process of γ- to β-globin. However, in silico analysis revealed no difference in the binding sites for rs10128556 with either allele suggesting that another variant in LD with this rs10128556 might be the functional one regulating the γ/β-globin genes (Hu *et al.* 2020). This work has not replicated in the sample of normal Portuguese the reported statistical significant association between rs10128556 and levels of HbF.

The BGLT3 gene emerged very recently as a positive regulator of the γ-globin genes (Ivaldi *et al.* 2018) but the relationship between genetic variations at BGLT3 and HbF levels is still unclear. As previously mentioned, Santos (2019) detected strong association between the rs7924684 SNP, located in the BGLT3 gene, and levels of HbF in Portuguese β-thalassemia minor patients. In contrast, the results obtain in the present study with Portuguese subjects lacking any hemoglobinopathy and showing normal levels of HbF did not showed up similar association.

The Kolmogorov-Smirnov normality test revealed that the HbF data was not normally distributed. Thus, the distribution of log-transformed HbF values according to SNP genotypes (genotypes homozygous for the ancestral allele versus homozygous and heterozygous for the derived allele) was analyzed by one-way-ANOVA using IBM® SPSS® Statistics v.26 software. Results are available in Table 15.

Gene	**SNP**	**Genotypes HbF Mean (SD)**		**P**
HBB	rs16911905	**GG** (n=45) 1.60 (1.73)	**GC + CC** (n=6) 2.07 (3.46)	0,895
Intergenic HBBP1-HBD	rs968857	**TT** (n=14) 1.86 (2.67)	**CT + CC** (n=37) 1,58 (1.65)	0.619
HBBP1	rs10128556	**GG** (n=20) 1.78 (1.61)	**GA + AA** (n=31) 1.57 (2.18)	0.182
BGLT3	rs7924684	**GG** (n=20) 1.96 (2.55)	**GA + AA** (n=31) 1.46 (1.48)	0.862
HBG2	**rs2070972**	**CC** (n=10) 2.35 (1.95)	**AC + AA** (n=24) 1.48 (1.95)	**0.042**
HBG2	rs113425530	**GG** (n=40) 1.38 (1.69)	**GT + TT** (n=11) 2.66 (2.57)	0.079
HBG2	rs7482144	**GG** (n=20) 1.78 (1.61)	**GA + AA** (n=31) 1.57 (2.18)	0.182

Table 15. Associations of the HbF levels (log-transformed) with genotypes of the studied SNPs in the 51 Portuguese individuals without hematological disorders. HbF mean levels shown are not-transformed.

n: number of subjects within each genotype (genotypes heterozygous and homozygous for the derived allele were combined). P: P-value obtained with the Kruskal-Wallis (1-way-ANOVA) test. Significant P-values (P<0.05) are in bold.

The analysis revealed a statistically significant association involving rs2070972 (P=0.042). The finding is in agreement with result obtained for this SNP in the case-control analysis under a recessive model (above presented), when a P-value below 0.05

was considered significant. Before Bonferroni correction for multiple testing, the case-control study under additive and dominant models also showed significant associations between the rs113425530 SNP and HbF levels. Although one-way-ANOVA failed to capture a significant correlation between this SNP and log-transformed HbF levels, the associated P-value of the test was once again rather low (P=0.079).

Figure 14 shows, in box-plots, the distribution of log-transformed HbF levels within each genotype of these two SNPs. Individuals who where homozygous for the minor ancestral allele (C) of the rs2070972 SNP showed higher values when compared to those with the derived allele (A). In contrast, subjects who where homozygous for the major ancestral allele (G) of the rs113425530 SNP showed slightly lower values when compared to the individuals with the derived allele (T). Intriguingly, the conventional HBB*S haplotypes correlated with high levels of HbF in SCD patients (AI and SEN) are defined by the alleles A and G of the SNPs rs2070972 and rs113425530, respectively, and the same occurs in the conventional haplotype Bantu/CAR that correlates with low levels of HbF. This observation underscores the complexity of interaction between sequences in the β-globin cluster that ultimatedely migh influence HbF levels.

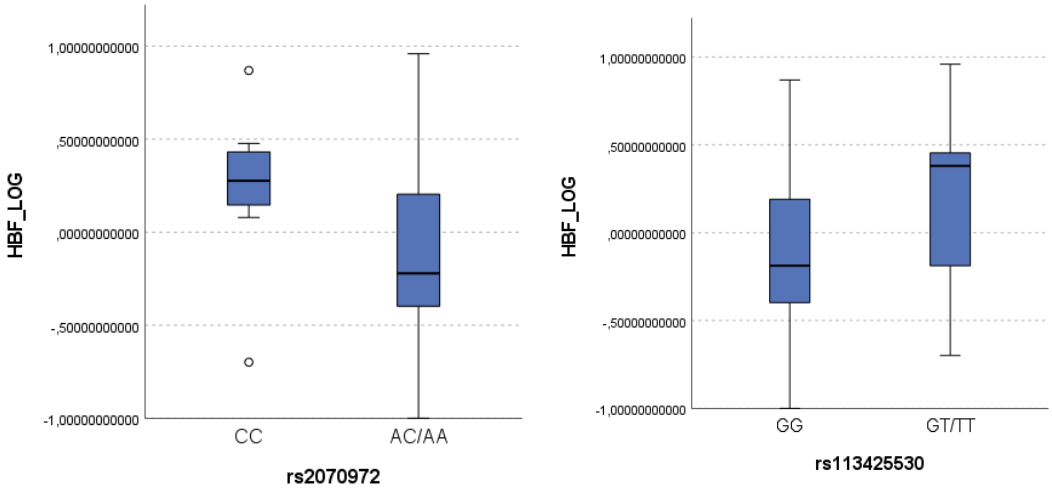


Figure 14. Box-plots showing the distribution of log-transformed HbF levels within genotypes of the SNPs rs2070972 and rs113425530 in Portuguese individuals without hematological disorders. Each rectangle represents the data between the 25th and 75th quartiles, and the bar within each rectangle is the median value for HbF.

When compared to other case-control studies conducted in normal individuals, the not fully congruent results here obtained after assessing the association between HbF levels at each locus in Portuguese subjects without hematological disorders raises questions that are difficult to answer for now. A serious limitation in our study is the reduced number of individuals analyzed. Another hindrance is the cutoff-point used to distinguish normal versus raised HbF levels in diferent case-control studies, which is often set at ~1%. These may be some of factors that apparently turn incoherent the results provided by different studies.

However, putting our results in the framework of other derived from studies that analyzed patients with SCD and/or β-thalassemia, including the one carried out by Santos (2019) in a sample of Portuguese carriers of β-thalassemia, it is likely to presume that the common polymorphisms at the β-globin clusters that more consistently showed evidence of being modulators of HbF levels in those patients, appear to have little or null influence in normal subjects. This indicates that physiological responses induced by the presence of β-hemoglobinopathies, including often stress erythropoiesis, might be crucial to activate regulatory mechanisms of HbF levels in which participate interactively a variety of sequences from the β- cluster, whose role in normal condictions is irrelevant or much more silent.

The last subject addressed in this work, was the association between HbF levels and multi-loci inferred haplotypes, which was prompted by the detected pattern of LD between the SNPs under study.

The results for the most common haplotypes detected based only on the 6 SNPs (rs7482144, rs113425530, rs2070972, rs10128556, rs968857 and rs16911905) defining the sickle cell haplotypes (Table 16) and the ones obtained when the rs7924684 in BGLT3 was also considered to infer 7 SNPs-haplotypes (Table 17) were very similar.

As previously mentioned, the conventional HBB*S haplotypes are well associated with a typical level of HbF in SCD patients: the SEN and AI haplotypes are correlated with high levels of HbF, the BEN and CAM haplotypes are associated with intermediate HbF levels and the Bantu/CAR haplotype is related with low levels of HbF in patients with SCD (Vinson *et al.* 2004).

Haplotype	Observations	HPFH (n=15)	Controls (n=36)	P
GGCGCG		0.44	0.33	0.304
AGAATG	Senegal	0.26	0.31	0.609
GTCGTG		0.14	0.01	**0.011**
GGAGTG	Cameroon	0	0.15	**0.024**

Table 16. Haplotype frequencies for the 6 SNPs that define the conventional HBB*S haplotypes and association analysis in individuals with without hematological disorders. Significant association P-values (P<0.05) are in bold. Haplotypes were as follows: rs7482144; rs113425530; rs2070972; rs10128556; rs968857; rs16911905.

The two most frequent haplotypes, one of them corresponding to the SEN haplotype (previously correlated with the highest levels of HbF) did not showed significant associations. However, the GTCGTG (or the GTCGGTG when considered the rs7924684) haplotype, which contains the two alleles at rs113425530 (T) and rs2070972 (C) found tending to raise HbF, was much more frequent (0.14) in HPFH individuals than in those with normal HbF values (0.01). Accordingly it revealed a significant association with HbF levels (P<0.05; P = 0.011). This haplotype did not matched any of the conventional HBB*S haplotypes, however only differed from the BEN haplotype (associated with intermediate HbF levels) in rs16911905, which was a SNP that did not revealed to affect HbF. On the other hand, the GGAGTG (or GGAGGTG) haplotype, which corresponds to the CAM haplotype (also related to intermediate HbF levels) and contains the complementary alleles of the rs113425530 (G) and rs2070972 (A) SNPs, was absent in the subjects classified as HPFH, while attained the frequency of 0.15 in the control group. This haplotype also showed a significant correlation with levels of HbF (P<0.05; P = 0.024).

Haplotype	Observations	HPFH (n=15)	Controls (n=36)	P
GGCAGCG		0.33	0.35	0.908
AGAGATG	Senegal + rs7924684 G allele	0.26	0.32	0.593
GTCGGTG		0.13	0.02	**0.015**
GGAGGTG	Cameroon + rs7924684 G allele	0	0.16	**0.019**

Table 17. Haplotype frequencies for the 7 SNPs and association analysis in individuals without hematological disorders. Significant association P-values (P<0.05) are in bold. Haplotypes were as follows: rs7482144; rs113425530; rs2070972; rs7924684; rs10128556; rs968857; rs16911905.

Out of the two late haplotypes, only the CAM was present in the sample of β-thalassemia carriers and was not found to be associated with HbF levels, contrarily to two other haplotypes, AGAATG (SEN) and GGAGCG, that revealed strong associations with HbF (Santos, 2019) .

These findings suggest once more that physiological conditions still poorly understood seem to be important to induce molecular strategies regulating levels of HbF that may imply different sequences in the β-cluster.

5. Conclusions and Future Perspectives

HbF is a main modulator of the clinical and hematological features of individuals affected with sickle cell anemia and β-thalassemia. Due to that, several therapeutic approaches are being designed to achieve an effective increase in HbF levels in patients with β-hemoglobinopathies, among which are those targeting the genetic modifiers of HbF.

This has prompted great interest in understanding which genetic polymorphisms influence HbF levels, and a few are now relatively well established, with some of them being trans and other cis-acting factors. Within the β-globin cluster, the most well-known polymorphism associated with HbF is rs7482144 (XmnI), whose role has been mainly documented in individuals presenting the pathophysiological patterns typical of sickle cell anemia or β-thalassemia. However, the influence on HbF of other genetic variations in the β-globin cluster still remains largely uncertain. Very recently evidence emerged that two non-coding regions in the cluster, BGLT3 and HBBP1, could also account for HbF. HBBP1 contains a SNP (rs10128556) that since long is interrogated to infer the background haplotypes of pathogenic mutations in the β-cluster. It is also known that some of those major haplotypes are associated with different levels of HbF.

In this context, with this work we sought to analyze whether variation at 7 SNPs (one in BGLT3, other in HBBP1 and the remaining spread throughout the β- cluster) were related with HbF in Portuguese subjects without hemoglobinopathies.

In a previous work (Santos, 2019), the same set of SNPs had been assessed in Portuguese carriers of β-thalassemia, allowing to detect very strong associations between levels of HbF and 4 of those SNPs: rs7482144, rs7924684, rs101288556 and rs968857.

In this study, however, no such strong associations were replicated. The studied subjects, encompassed individuals with either normal or high levels HbF. When a case control analysis was conducted, only the SNPs rs113425530 and rs2070972, located in an intronic region of HBG2, revealed a significant P-value in the association tests performed under different models, which nevertheless would lose significance when correction for multiple testing was applied. Until now, such polymorphism per se has never been reported to regulate HbF levels in normal individuals, and in the sample of β-thalassemia carriers analyzed by Santos it also didn't appear to influence HbF. The signal here obtained for this SNP is faint and might have been fortuitous, given the small size of the sample. Still, it deserves attention and further investigation.

Among the sub-sample of subjects with normal HbF levels, the association analysis also revealed a marginal significance in a SNP, this time rs968857, located in the intergenic region between HBBP1 and HBD. This was one of the SNPs strongly associated with HbF among the Portuguese β-thalassemia carriers. Contrarily, the marginal significant association here detected in normal subjects did not resist the correction for multiple tests.

In short, our results contrasted overtly with those obtained with β-thalassemia carriers, and also failed to replicate associations previous reported in normal subjects implicating some of the SNPs from the β-cluster here scrutinized.

A major limitation of our work was the low number of subjects analyzed, and so, it will be necessary to enlarge that number in order to increase the power of the study.

Nevertheless, one important message brought by this study with the contrasting findings in normal and β-thalassemia carriers of Portuguese origin, is that the physiological conditions, some possibly associated with stress erythropoiesis, appear to be critical to trigger molecular mechanisms in which participate different sequences of the β-cluster able to induce HbF. Eventually, variability on these sequences might only exert a role as modulator of HbF in specific conditions.

In the future, it would be interesting to analyze with this same set of SNPs a sample of non-transfusion-dependent β-thalassemia intermedia subjects in order to obtain a better pathophysiological scenario to address the relationship between those SNPs and levels of HbF.

6. Bibliographic References

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.

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.

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.

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.

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

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.

Dabke, P., Colah, R., Ghosh, K., & Nadkarni, A. (2013). Effect of Cis Acting Potential Regulators in the β Globin Gene Cluster on the Production of HbF in Thalassemia Patients. *Mediterranean journal of hematology and infectious diseases,* 5(1), e2013012.

de Souza Carrocini, G. C., Venancio, L. P., & Bonini-Domingos, C. R. (2015). Screening of Transcription Factors Involved in Fetal Hemoglobin Regulation Using Phylogenetic Footprinting. *Evolutionary bioinformatics online*, 11, 239–244.

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.

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

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.

Guida, V., Cappabianca, M. P., Colosimo, A., Rafanelli, F., Amato, A., & Dallapiccola, B. (2006). Influence of Ggamma-158C --> and beta- (AT)x(T)y globin gene polymorphisms on HbF levels in Italian beta-thalassemia carriers and wild-type subjects. *Haematologica,* 91(9), 1275–1276.

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.

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.

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 threedimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. *Genes and Development.* 31(16):1704-1713.

Hu, Li & Huang, Ling & Han, Yuanyuan & Jin, Tingting & Liu, Juan & Jiang, Minmin & Liu, Xingmei & Li, Yuanyuan & Han, Wenping & An, Bangquan & Huang, Shengwen. (2020). Association of polymorphisms in the HBG1-HBD intergenic region with HbF levels. *Journal of Clinical Laboratory Analysis.* 34.

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.

Jouini, Latifa & Bibi, Amina & Ouali, Faida & hadj fredj, Sondess & Ouennich, Fekria & Hajer, Siala & Messaoud, Taieb & Fattoum, Slaheddine. (2011). Contribution of β-globin

cluster polymorphisms to raise fetal hemoglobin levels in normal adults. *Molecular biology reports.* 39. 4619-25.

Kaushansky, K., Lichtman, M.A., Beutler, E., Kipps, T.J., Seligsohn, U., and Prchal, J.T. (2010). Williams Hematology, Eighth edn (New York: McGraw-Hill Medical).

Kiefer, 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.

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 Gγ-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.

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 sicklecell anemia in Triângulo Mineiro, Minas Gerais, Brazil. *Jornal Brasileiro de Patologia e Medicina Laboratorial.* 52(1):6-10.

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.

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.

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

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.

Neumayr, L. D., Hoppe, C. C., & Brown, C. (2019). Sickle cell disease: current treatment and emerging therapies. *The American journal of managed care,* 25(18 Suppl), S335–S343.

Nuinoon M.; Makarasara W.; Mushiroda T et al. (2010). A genome-wide association identified the common genetic variants influence disease severity in β0 - thalassemia/hemoglobin E.

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.

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

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.

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

Santos R. (2019). Genetic variations in the β-globin cluster: influence on levels of fetal hemoglobin (Master's thesis). Retrieved from the Open Repository of Universidade do Porto.

Saraf, S. L., & Rondelli, D. (2019). Allogeneic Hematopoietic Stem Cell Transplantation for Adults with Sickle Cell Disease. *Journal of clinical medicine,* 8(10), 1565.

Sherva, R., Sripichai, O., Abel, K., Ma, Q., Whitacre, J., Angkachatchai, V., Makarasara, W., Winichagoon, P., Svasti, S., Fucharoen, S., Braun, A., & Farrer, L. A. (2010). Genetic modifiers of Hb E/beta0 thalassemia identified by a two-stage genome-wide association study. *BMC medical genetics*, 11, 51.

Siatecka, M., & Bieker, J. J. (2011). The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood,* 118(8), 2044–2054.

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.

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

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 βS -Globin Haplotypes: Sequencing of the Aγ-IVS-II Region. *Hemoglobin.* 28(4):317-323.

Waye, John & Chui, D.H.K.. (2001). The α-globin gene cluster: Genetics and disorders. Clinical and investigative medicine. *Médecine clinique et experimentale*. 24. 103-9.

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.

Zertal-Zidani, Samia & Ducrocq, Rolande & Sahbatou, Mourad & Satta, Dalila & Krishnamoorthy, Rajagopal. (2002). Foetal haemoglobin in normal healthy adults: Relationship with polymorphic sequences cis to the β globin gene. *European journal of human genetics* : EJHG. 10. 320-6.

Supplementary Data

Sample	Polymorphisms						
	rs7482144	rs113425530	rs2070972	rs7924684	rs10128556	rs968857	rs16911905
2	GA	GG	AC	GA	GA	CT	GC
27	AA	GG	AA	GG	AA	TT	GG
132	GG	GG	AC	AA	GG	CC	GG
147	GG	GG	AA	GA	GG	CT	GG
223	GG	GG	AC	GA	GG	CT	GG
247	GG	GG	AC	GA	GG	CT	GG
294	AA	GG	AA	GG	AA	TT	GG
313	GA	GG	AC	GA	GA	CT	GG
415	GG	GG	CC	AA	GG	CC	GG
427	GG	GG	AC	GA	GG	CT	GG
629	GG	GT	CC	GA	GG	CT	GG
649	GA	GT	AA	GG	GA	TT	GG
672	GG	GG	CC	AA	GG	CC	GG
774	GA	GG	AC	GA	GA	CC	GG
781	GG	GT	CC	GA	GG	CT	GG
783	GA	GG	AC	GA	GA	CT	GG
823	GA	GG	AC	GA	GA	CT	GG
829	GA	GG	AC	GA	GA	CT	GG
830	GA	GG	CC	GA	GA	CT	GG
845	GG	GG	AC	GA	GG	CT	GG
867	GA	GG	AA	GG	GA	TT	GG
876	GG	GG	AC	GA	GG	CT	GC
912	GG	GG	AC	GA	GG	CT	GG
914	GG	GG	AC	GA	GG	CT	GG
922	GA	GG	AC	GG	GA	CT	GG
945	GA	GG	AA	GG	GA	TT	GC
994	GA	GG	AA	GG	GA	TT	GG
2032	GA	GG	AC	GA	GA	CT	GG
2493	GA	GG	AC	GA	GA	CT	CC
2750	GA	GT	AA	GG	GA	TT	GC
2751	GA	GT	AA	GG	AA	TT	GG
2779	GG	GG	AC	GA	GG	CT	GG
3332	GG	GG	CC	AA	GG	CC	GG
3383	GA	GG	AC	GA	GA	CT	GG
3939	GA	GG	AA	GG	AA	TT	GG
4608	GA	GG	AA	GG	AA	TT	GG
4812	GG	GT	CC	GA	GG	CT	GG
6308	GA	GG	AC	GA	GA	CT	GG
6326	GA	GG	AC	GG	GA	CT	GG
6903	GA	GT	AA	GG	GA	TT	GG
6939	GG	GG	CC	AA	GG	CC	GG
6960	GA	GG	AA	GG	AA	TT	GG
6996	GG	GT	CC	GG	GG	CT	GG
7000	AA	GG	AA	GG	AA	TT	GG
7033	GA	GG	AA	GG	GA	CT	GG
7046	GA	GT	AA	GG	GA	TT	GG
7503	GA	GG	AC	GA	GA	CT	GG
8033	GA	GG	AC	GG	GA	CT	GG

8519	GG	GT	CC	GA	GG	CT	GG
8533	GA	GG	AC	GA	GA	CT	GG
9243	GG	TT	AA	GG	GG	CT	GC

Table S1 - Genotypic data of the polymorphisms found in the 51 studied samples from Portuguese individuals with normal hematological parameters.