UNIVERSITÀ DEGLI STUDI DI CAGLIARI

FACOLTÀ DI MEDICINA E CHIRURGIA

Dipartimento di Scienze Biomediche e Biotecnologiche

Dottorato di Ricerca in

Terapia pediatrica e Farmacologia dello Sviluppo (XXIII ciclo)

"STATISTICAL GENETICS APPLIED TO β-THALASSEMIA PHENOTYPE SEVERITY"

Coordinatore Scientifico: Prof. Renzo Galanello

Tutor: Prof. Renzo Galanello **Tesi di dottorato del Dott. Fabrice Danjou**

Anno Accademico 2009-2010

STATISTICAL GENETICS APPLIED TO β-THALASSEMIA PHENOTYPE SEVERITY

To Paola, Ernest, and the one to come…

ACKNOWLEDGEMENTS

This thesis is part of the research that has been done since I came to Prof. Galanello's team. It is therefore the result of a continuous exchange with all collaborators from the Molecular Genetic Laboratory of the *Ospedale Microcitemico di Cagliari*, to whom I would like to record my gratitude. Specially, I am glad to have the opportunity to acknowledge Dr. Franco Anni for his crucial contribution to this and other researches we did together, for his everyday help in any form, his patience and friendship.

This thesis would not have been possible without the support from the Department of Genetics of the Abramson Research Center of the Children Hospital of Philadelphia, and I would like here to acknowledge the numerous persons who helped me along the way. In particular, it is a pleasure to gratefully thank Prof. Marcella Devoto for her precious support: she gave me invaluable insights and suggestions from the first moments of this research.

Obviously, I also need to thank Paola without whom this work, and myself, would not have existed.

A special thought is devoted to my parents, brother and sister for their never-ending support.

Finally, I would like to thank all the patients of the *Ospedale Microcitemico di Cagliari*, because this work is not only for them but also their.

INDEX

TABLES

FIGURES

INTRODUCTION

Hemoglobinopathies are among the most common monogenic disorders in the world and it is estimated that about 1.5% of the global population (80 to 90 million people) are carriers of β-thalassemia. The total annual incidence of symptomatic individuals is estimated at 1 in 100,000 throughout the world and 1 in 10,000 people in the European Union. [1]

β-thalassemia is prevalent in populations in the Mediterranean, Middle East, Transcaucasus, Central Asia, Indian subcontinent, and Far East. It is also common in populations of African heritage. The highest incidences are reported in Cyprus (14%), Sardinia (10.3%), and Southeast Asia. The frequencies of β-thalassemia are particularly high in the malarial tropical and sub-tropical regions of Asia, Mediterranean and the Middle East, because of a selective advantage of heterozygotes against the severe forms of malaria.[2] However, because of population migration β-thalassemia is now also common in northern Europe, North and South America, the Caribbean, and Australia.[3]

Nowadays, in the less developed countries, affected children are on the increase due to falling childhood mortality from improved nutrition and better infection control, while in the more developed countries, epidemiology of the disease has been affected by a fall in total birth rate and preventive programs.

Carriers (heterozygotes) of β-thalassemia are clinically normal and largely unaware of their genetic condition. Inheritance of two copies of β-thalassemia genes, one from each parent, usually results in life-threatening anemia and transfusion-dependence for survival. Intermediate clinical forms of the disease exist as thalassemia intermedia, which is characterized by moderately severe anemia with the occasional need for blood transfusion. The mainstay of treatment for the severe forms (thalassemia major) is blood transfusion and aggressive management of complications including life-long iron chelation therapy.

The complications of iron overload, together with the sequels of the anemia and ineffective erythropoiesis, and the chelation therapy itself, are major causes of morbidity and mortality in the transfusion dependent β-thalassemias.[4] Although, regular blood transfusions in combination with aggressive iron chelation have been remarkably effective in delaying the onset of iron-related organ failure and improving mortality, many patients continue to be affected by cardiac disease, delayed pubertal maturation, and develop endocrine failure.[5] Further, the commitment to this rigorous life-long treatment regime is onerous and an enormous imposition on quality of life.

Hence, the continuing quest is for a molecular-based strategy, which, at best, will lead to a cure, and, at the very least, can ameliorate the anemia in the severe transfusion-dependent β-thalassemias. To this regard recent advances in the understanding of γ gene expression mechanisms bring hope. However, much has still to be done to better define the overall molecular basis of β-thalassemia. Such knowledge could not only bring to a molecular-based strategy but also improve the prediction of the phenotype from genotype in the prevention and management of β-thalassemia or even other genetic diseases.

BACKGROUND

β-thalassemias (OMIM 141900) are a heterogeneous group of autosomal recessive disorders characterized by reduced (β^+) or absent (β^0) production of adult hemoglobin A ($\alpha_2\beta_2$) β-globin chains. The main pathophysiologic mechanism is the imbalance between α and non α-globin chains: the non assembled α-globin chains precipitate in the form of inclusions. These α-globin chain inclusions damage the erythroid precursors in the bone marrow and in the spleen, causing ineffective erythropoiesis. More than 200 mutations of different severity have been described in the β-globin gene.[6-8] The wide variability of globin chain imbalance in homozygous β-thalassemia result in a wide spectrum of clinical manifestations, ranging from thalassemia major to thalassemia intermedia. Thalassemia major patients have a very severe anemia from early infancy, which needs lifelong blood transfusions. On the other hand thalassemia intermedia patients have a less severe anemia and survive without transfusions or with sporadic transfusions.

Since the key factor for the severity of β-thalassemia is the imbalance between α and non α-globin chains, any determinant able to reduce this imbalance is expected to produce a milder clinical phenotype. Mechanisms so far identified as responsible of thalassemia intermedia are:

a) presence of mild or silent β^+-thalassemia mutations with consistent residual production of β-globin chains (and hence HbA), and reduced excess of free α chains. Mild or silent β^+-thalassemia mutations are some transcriptional mutants (i.e. –101 C→T; -87 C→G; -30 T→A), some mutations activating alternate splice sites (i.e. cd 19 A→G; cd 24 T→A), some mutations in the consensus sequences (i.e. IVS 1-6 T→C; IVS2-844 C→G) and polyadenilation mutations;

b) co-inherited α thalassemia defects, able to consistently reduce the α-globin chain excess present in β-thalassemia patients. The ameliorating clinical effect of co-inherited α thalassemia is more predictable in presence of β^+-thalassemia mutations, while in homozygous β^0-thalassemia patients producing only HbF, even the coinheritance of two α-globin gene deletion not always results in a mild phenotype;

c) co-transmitted determinants increasing the γ chain production and hence HbF ($\alpha_2\gamma_2$). These determinants may be related to the nature of the thalassemia mutation itself, such as $\delta\beta^0$-thalassemias caused by deletions of variable size in the β-globin cluster and some rare β^0-thalassemia deletions extending into the β-globin promoter. Single point mutations of the hemoglobin Gγ (HBG2) or hemoglobin Aγ (HBG1) gene promoter can also be responsible of increased HbF production, the most common being a single-base substitution C to T at position 158 upstream of the transcription start site of the HBG2 gene, which is silent in normal individuals and in β-thalassemia heterozygotes, but leads to increased HbF production in individuals with erythropoietic stress, as occurs in homozygous β-thalassemia. This HBG2 mutation, sometimes referred to as Gγ-158 C>T is in linkage disequilibrium (in cis configuration) with some β-globin gene mutations.

Co-inherited determinants increasing HbF may or may not be linked to the β-globin gene cluster, such as the non deletion hereditary persistence of fetal hemoglobin. Recent studies using genome-wide association studies (GWAS) have identified quantitative trait loci (QTLs) independent from the β-globin gene cluster and it is likely that many other HbF-associated QTLs also exist.

In humans, a shift from γ to β-globin gene expression around birth, underlies the switch from fetal to adult hemoglobin production such that by six months of age, the major hemoglobin is HbA ($\alpha_2\beta_2$).[9] However, residual amounts of HbF continue to be synthesized throughout adult life and the continuous distribution of HbF trait suggests it is a quantitative trait, accounting in adults for <1% of total hemoglobin.[10] Its persistence in adult life is highly variable and genetically controlled with a heritability of 0.89.[11] Pancellular forms of hereditary persistence of fetal hemoglobin (HPFH) are relatively rare, while surveys of HbF distribution in normal individuals suggest that heterocellular

forms (hHPFH), defined as HbF levels between 0.8% and 5%, may be present in about 10% of the population.[12,13]

All β-thalassemias have variable increases in HbF, but one extreme form can be observed in β°-thalassemia intermedia patients who are transfusion independent with a mild disease despite the absence of HbA.[14-17] In the majority of β-thalassemia a large part of the HbF response is related to the erythropoietic stress, expanded erythroid mass, and preferential survival of the red cell precursors that contain HbF. Still, as previously mentioned, studies on the interaction of hHPFH with β-thalassemia suggested that the high HbF determinant can be linked to the β-globin locus, or can segregate independently implicating trans-acting QTLs.[14,18-23]

To date, three major loci that modulate HbF levels have been identified – Xmn1-HBG2 on 11p15.4, HBS1L-MYB intergenic region on 6q23.3, and BCL11A on 2p16.1 – contributing to 20–50% of the trait variance in patients with β-thalassemia, sickle cell disease and in healthy European Caucasians.[24-42] In the HBS1L-MYB locus it seems that MYB is causally involved through alteration of erythroid kinetics, while BCL11A has shown to act together with SOX6 via direct transcriptional repression of the γ-globin genes and to be activated by KLF1, whereas GATA-1 has shown to be a widespread and general regulator of transcription in erythroid cells that binds to α and β-globin as well as to HBS1L-MYB and BCL11A loci.[25,43-48] However, evidence exist for other contributing loci as the 8q and Xp22.2–22.3 QTLs, but were not validated in recent GWAS that mainly studied moderately raised levels of HbF.[49-53]

In addition, other determinants than HbF levels might contribute to the phenotypic severity of β-thalassemia, through the coinheritance of other genetic factors mapping outside the β-globin gene cluster (the best known example being UGT1A: its mutation causing Gilbert disease leads to increased jaundice and increased risk of gallstones in thalassemic patients).[54] A part of the still unexplained phenotypic variation probably involves the proteolytic capacity of the erythroid precursors in catabolizing the excess α-globin chains and the destruction mechanisms of red cell progenitors and their progeny, including mechanical damage, interference with cell division and oxidative destruction of both organelles and components of the red cell membrane.[55] The nitric oxide depletion from chronic hemolysis together with abnormal red blood cell (RBC) phospholipid membrane asymmetry might also be related to the pathogenesis of hypercoagulability, a chronic condition of thalassemic patients that also involves platelets, endothelial cells, monocytes and peripheral blood activation.[56,57]

In summary, while the genetic determinants of β⁺-thalassemia phenotypic variations have been largely defined, the inherited modifying factors able to ameliorate the clinical features in the majority of patients with β°-thalassemiahave been only partially clarified yet. In fact, the presence of mild or silent β⁺-thalassemia defects, with a relatively high residual output of β-globin synthesis, is adequate to produce a mild clinical picture. The coinheritance in these patients of α thalassemia, additionally reducing the α/non-α imbalance, further ameliorates the clinical phenotype. On the other hand, the molecular basis of β°-thalassemia has greatly benefited from recent studies on genetic determinants of HbF levels, but the variation of disease severity is still not fully understood.

AIM, RATIONAL AND HYPOTHESIS

AIM

The goal of this study is to identify and characterize new candidate modifier genes which are effective in ameliorating the clinical phenotype of β°-thalassemia patients. Knowledge of the control of developmental expression of these genes may lead to targeting specific modifiers in order to correct or moderate the effect of α/non-α chains imbalance, to decrease clinical severity in patients with β-thalassemia and sickle-cell anemia.

RATIONALE

All subjects included in the study are homozygous for the β39 C->T mutation, with no residual production of β chains and all negatives for the Gγ-158 C>T mutation. Still, patients with thalassemia major are lifelong dependent on blood transfusion (transfusion dependent – TD) while patients with thalassemia intermedia can survive without transfusion (non transfusion dependent – NTD). In this last category of patients, the amount of γ chains necessary to reduce the α chains excess (and the ineffective erythropoiesis) through formation of $\alpha_2\,\gamma_2$ tetramers (HbF) should be elevated. These patients are able to maintain a hemoglobin levels of 7 to 11 g/dl without red blood cell transfusions and their hemoglobin is composed of only HbF and traces of HbA_2.

Therefore, some determinants exist that are responsible for the difference between TD and NTD patients and should strongly affect HbF levels. However differences, mostly within TD group, make this separation less evident: some TD patients might start transfusion at birth whereas others might start blood transfusions after years. Furthermore, for some TD patients, decision to start transfusion could have been required by temporary environmental factors but conduced to transfusion dependency and TD status. Hence it could be more appropriate to study the time to first transfusion than the TD status for two reasons: first because it should be more accurate, defining both TD/NTD group membership and within groups differences, and second because it should allow an analysis of other factors than HbF levels also responsible for phenotype amelioration.

The considered population represents a unique model to study genes ameliorating the severe phenotype of homozygous β°-thalassemia, and eventually uncover loci contributing to HPFH for which evidences exist but were not yet validated in recent GWAS, as the 8q and Xp22.2–22.3 QTLs.[49-53] In particular, the linkage to the chromosome Xp22.2–p22.3 locus was identified in a population associated with the inheritance of an extreme high fetal cells trait and early studies of families with sickle cell disease suggested high HbF determinants that segregated independently of the β-globin locus.[58] It is therefore conceivable that this finding represent linkage to a form of HPFH not represented in subsequent studies to date, as recent GWAS investigated levels of moderately "raised" HbF or with microarrays scarcely covering the X chromosome.[24,59-62] Besides HbF levels, many other factors responsible for amelioration for the phenotype of homozygous β°-thalassemia are still to be discovered, as any element inducing better survival of red blood cells, normally making no difference to healthy subjects, could allow β°-thalassemia patients to start later or never start transfusions.

We therefore propose in the present study to perform a whole genome scan using hybrid technology to test for association of time to first transfusion with SNPs and association of TD status with rare CNVs, linked to genes or gene pathways that contribute to severity of homozygous β°-thalassemia phenotype. We used the Affymetrix Gene Chip Mapping 6.0 microarray, an assay that enables highly representative genome-wide coverage of SNPs and extensive coverage of known CNVs and polymorphic regions of the genome. We took advantage of the existence of a unique and homogeneous group of 386 Sardinian patients, all homozygous for the β39 C->T mutation and followed at the

Ospedale Microcitemico di Cagliari, to try to uncover independent SNPs and rare CNVs representing genes or candidate regions associated with the modulation of homozygous β°-thalassemia phenotype.

HYPOTHESIS

The main hypothesis assumed in the present study are the following:

- Time to transfusion is a phenotype able to reflect correctly the underlying endophenotypes that are HbF levels and other elements eventual contributing to the variation of time to first transfusion.
- TD versus NTD is a phenotype able to reflect correctly the underlying endophenotypes that are HbF levels and other elements eventual contributing to the variation of time to first transfusion.
- Some loci influencing β°-thalassemia phenotype are yet to discover.
- SNPs and/or CNVs are relevant to identify such loci.
- CNVs influencing TD status are rare.
- The effects of such loci are important enough.

Regarding the last point, according to the power estimation used to plan our study, under the additive model, we have ≥ 70% power for detecting variants with genetic relative risk ≥ 4 over a wide range of allele frequencies (MAF ≥ 0.19), down to a genetic relative risk of 3.4 for a MAF of 41% . The estimated power was calculated for a GWAS of 59 cases and 327 controls with an overall p-value of 1.00E-07 (corresponding to a genome-wide significance level of 0.05 after a stringent Bonferroni correction for multiple testing of 500,000 independent SNPs) using Quanto software.[63] This hypothesis is the most stringent that we used (gene based joint testing SNP analysis for TD status, see results), however the power of the present analysis was increased by our choice to use survival analysis for SNP analysis. Regarding CNV analysis, we reduced greatly the burden of markers using only consensual CNVs among different algorithms.

METHODOLOGY

PATIENTS

The following patients were studied: 59 patients with β°-thalassemia intermedia and 327 with thalassemia major with the same β-globin gene mutation (homozygotes for the β39 C->T mutation). This population is composed of 386 founders from Sardinia, all negative for the Gγ-158 C>T mutation, divided in 201 males and 185 females.

In Sardinia about 10% of patients homozygotes for β° 39 C→T non-sense mutation, have the mild clinical phenotype of β°-thalassemia intermedia. These patients are able to maintain a hemoglobin levels of 7 to 11 g/dl without red blood cell transfusions (i.e.: non transfusion dependent patients - NTD) and their hemoglobin is composed of only HbF and traces of HbA_2. This population of patients represents a unique model for studying the effect of genes able to ameliorate the severe clinical phenotype of homozygous β°-thalassemia.

All patients are from Sardinia, an island at least geographically "distinct" and therefore less susceptible to genetic admixing. Consequently, influx of different mutations on different haplotype backgrounds causing the same phenotype should be minimized, making this population ideal for the study.

For each patient the following information were available:

- gender;
- age at first transfusion or age at last follow-up;
- β39 C->T mutation status (all patients are homozygotes for such mutation);
- co-inherited α thalassemia defects.

Table 1. Characteristics of the studied population.

Gender (Number of cases)	Males	201
	Females	185
Alpha thalassemia defects (Number of cases)	-α/-α	41
	-α/αHphIα	1
	-α/αNcoIα	5
	-α3,7/-α4,2	2
	-α/αα	113
	-α4,2/αα	2
	αHphIα/αα	2
	αNcoIα/αα	13
	-α/ααα	1
	αα/αα	204
	ααα/αα	2
Age at first transfusion [TD patients]	Median = 8 months 26 days IQR = 1 year 3 months	
Age at last follow-up [NTD patients]	Mean = 41 years 3 months Std. Dev. = 12 years	

All participants enrolled were volunteers who signed an informed consent after detailed explanations on the purpose and modalities of the study project.

PROTOCOL

The study consisted in a whole genome scan on 386 samples to test for association between time to first transfusion and SNPs and between TD status and rare CNVs, to find genes or candidate regions that contribute to the amelioration of homozygous β°-thalassemia phenotype.

We performed such hybrid technology GWAS using the Affymetrix Gene Chip Mapping 6.0 microarray, that consists of about 906,600 SNPs sequences and about 900,000 non polymorphic oligonucleotides, covering the whole genome with an average spacing of 0.7 Kb (Affymetrix, Santa Clara, CA, USA). To such purpose we digested genomic DNA (250 ng) in 2 separate reactions with *Nsp I* or *Sty I* (New England Biolabs, Ipswich, MA, USA), as recommended by the manufacturer (Affymetrix). Following digestion, an adaptor was linked to the restricted fragments, the reaction diluted 4X and the fragments amplified by PCR. After purification using Magnetic Beads (Agencourt Bioscience Corporation, Beverly, MA, USA), 90 µg of PCR products were fragmented and end labeled using 30 U/µl of terminal deoxynucleotidyl transferase, and then hybridized for 16-18 hours to the Affymetrix 6.0 chip at 49^0 C in the Gene Chip Hybridization Oven 640 (Affymetrix). Chips were washed, stained in the GeneChip Fluidic Station 450 and scanned with the Scanner 3000 7G (Affymetrix).

Successively Affymetrix CEL files were loaded into the proprietary software Genotyping Console (GTC) v. 4.0 (Affymetrix) for genotype and copy number analysis. The Birdseed v. 2.0 genotype calling algorithm was used to perform a multiple-chip analysis to estimate a signal intensity for each allele of each SNP, fitting probe-specific effects to increase precision. It then made genotype calls by fitting a Gaussian mixture model in the two-dimensional A signal versus B signal space, using SNP specific models to improve accuracy. Copy number analysis was performed with GTC v. 4.0 using the Affymetrix HapMap270 reference model file for comparison, while segmentation was also performed using two other algorithms (PENNCNV and QUANTISNP) to obtain a pool of consensual segments from which we analyzed rare CNVs.[64-67]

Genome-wide analysis of SNP frequencies were implemented using PLINK v1.07 and R v. 2.9.2, after imputation on CEPH (Utah residents with ancestry from northern and western Europe - CEU) plus Tuscany (Italians from Toscana - TSI) samples from The International HapMap Consortium, using MACH v.1.0.16 software.[68-71] Prior to imputation, quality control procedures were implemented to remove "failed calls markers". We excluded samples with low call rates, samples from related subjects or indicating eventual contamination or replication through identity by state analysis, and samples with discrepancies between gender and computed sex. We also checked for samples with inbreeding coefficient indicating possible contamination or outliers at nearest neighbor analysis and for outliers at principal component analysis (PCA) using the software package EIGENSTRAT.[72] We flagged SNPs for low genotype call rate, low minor allele frequency and failure of Hardy-Weinberg equilibrium. Association with time to first transfusion was done through survival analysis using a Cox proportional hazards model. SNPs with association p-values less than 1.00E-07 were considered statistically significant.

For CNV analysis, intensity signals of samples were treated using GTC v. 4.0 to produce quality control indexes specific to CNV analysis process. We filtered out samples improper for analysis and corrected signals using the median of the absolute values of all pairwise difference (MAPD), regional GC correction and the number of CNVs per sample. Then, we performed copy number segments analysis using GTC v. 4.0 as well as two other segmentation algorithms (PENNCNV and QUANTISNP) to select CNVs found in at least two of the three different segmentation algorithms. Settings were identical for all three algorithms, considering a minimum of 5 markers showing consensus for gain or loss spanning at least a 100 kb region. To consider only rare CNVs, all CNVs present in more than 1% of TD samples were eliminated from further analysis. We then ran analysis on the remaining CNVs to check association with TD status: case/control analysis for association with the phenotype was done using

permutation procedures made available in PLINK v. 1.07.[73] Each rare copy number containing known genes or small RNA sequences resulting significant at genome-wide level were considered.

RESULTS

QUALITY CONTROL

CNVs

Copy number analysis, using the Affymetrix HapMap270 reference model file for comparison, was performed using GTC v. 4.0 (Affymetrix). The median of the absolute values of all pairwise difference (MAPD) was used to evaluate whether the chip produced data useful for copy number analysis. MAPD is defined as the median of the absolute values of all pairwise differences between log2 ratios for a given chip and Affymetrix recommends that the MAPD value be less than 0.4. In addition, we removed samples outliers for the number of CNVs.

Upon MAPD analysis and number of CNVs per sample, we removed 60 samples (11 NTD and 49 TD) for a remaining number of 326 samples (48 NTD and 278 TD).

An important part of quality control for CNVs also relayed on the GC correction we applied and on the filtering of consensual CNVs from different segmentation algorithms that we performed. These techniques are detailed in the following sections.

SNPs

We excluded samples with low sample call rates (<97%) and discrepancies between referred and calculated sex. We also analyzed samples for their identity by state using GRR software to eliminate samples from related subjects or with eventual contamination or replication (see Figure 1).[74] Furthermore, we developed a composite index for samples with marginal call rate (between 97 and 98.5%): for such sample we also considered the inbreeding coefficient (F) and the nearest neighbor score (Z) calculated with PLINK. We determined outlier samples for these two statistics and weighted samples call rate with their rank to obtain a composite index of these three measures, and eventually discard samples with marginal call rate and poor F or Z statistics. Finally we checked for outliers at principal components analysis (see next section) to eventually remove them, but no sample was above 6 standard deviations. Table 2 and Table 3 summarize quality control procedures on samples and population characteristics.

We also flagged SNPs for low SNP genotype call rate (< 98%), discrepancies in duplicate genotyping, or failure of Hardy-Weinberg equilibrium (HWE) (p-value <0.005). HWE test was conducted separately in NTD and TD patients using PEDSTATS.[75] While deviation from HWE in cases may indicate association and, therefore, could help in the identification of disease variants, deviation from HWE in controls is often assumed to derive from genotyping errors, non random pattern of missing data, or unrecognized population stratification, so that we tested for HWE among control subjects for genotyping quality control. We also eliminated SNPs with minor allele frequency below 1% and SNPs with heterozygous genotypes while on haploid chromosomes. After quality control, overall genotype rate was: 0.997 for a total of 647,632 markers (see Table 4).

Table 2. Quality control procedures on samples.

Quality control procedure	Before QC	Removed	Left for analysis
Call rate + inbreeding coefficient + nearest neighbor score	386	15	345
Sex discrepancies		11	
Identity by state		15 (8 twins/repeated + 7 sibs/contaminated)	

Table 3. Characteristics of population before and after quality control.

	Males	Females	NTD	TD
Before QC	201	184	59	327
Left for analysis	178	167	48	297

Table 4. Quality control procedures on SNPs.

Quality control procedure	Before QC	Removed	Left for analysis
Call rate <0.98	905,384	138,637	647,632
Minor allele frequency < 0.01		119,245	
Test for Hardy Weinberg equilibrium, p-value < 0.005		8,195	
Heterozygous haploid genotypes		21,703	

SEGMENTATION

Segmentation is maybe the most important step of CNV analysis based on chips. However great variability is known to result from different segmentation algorithms.[76] We therefore decided to concentrate on consensual CNVs using three algorithms, to reduce overall burden of CNVs and minimize false positive while maximizing true positives, based on the hypothesis that a CNV found by more than one algorithm has more chance to be reliable. We used, to this purpose, three software: the proprietary GTC v. 4.0 software (Affymetrix), PENNCNV and QUANTISNP.[64-67]

After quality control (see before) and filtering of CNVs present in less than 1% of TD patients, we segmented intensity signal with a default setting of at least 5 markers showing consensus for gain or loss spanning at least a 100 kb region for all three algorithms. The burden of CNV found by each algorithm was quite different (9,490 for GTC, 3,934 for PENNCNV and 6,507 for QUANTISNP) so we decided to consider CNVs consensual in at least two of the three algorithms, not to give more importance to the algorithm finding less segments (in the case we would have chosen to only consider CNVs consensual in all three algorithms).

To characterize consensual CNVs we used data management facilities for CNVs from PLINK v. 1.07 as well as ad-hoc UNIX syntax scripts and Excel spreadsheets (Microsoft Excel 2010, Microsoft, Redmond, Washington, USA).[68] We first converted CNVs status in only one deletion and one duplication status and then collapsed CNVs from different algorithms together if they shared at least one probe, giving to the resulting CNV its limits from the minimum and maximum of both CNVs extremities. We decided not to consider different the CNVs overlapping for less than a certain threshold to be able to also include complex CNVs (i.e.: with borders not well defined between samples) in further analysis and because it represented a conservative approach, only adding more CNVs to the global burden of consensual CNVs we kept .

We ascertained from such process that consensus between segmentation algorithms is low: GTC agreed with PENNCNV and QUANTISNP for 28% and 21% respectively (i.e.: 28% of CNVs found by GTC were also present in PENNCNV segmentation results), PENNCNV agreed with GTC and QUANTISNP for 68% and 39% respectively, and QUANTISNP agreed with GTC and PENNCNV for 30% and 24% respectively. PENNCNV specifically seems more sensitive but less specific, as its concordance rate is always higher but characterizing 2.5 and 1.7 times less CNVs than GTC and QUANTISNP respectively. However, even if concordance is scarce, among CNVs common between different segmentation algorithm a relatively high rate of physical overlap was found: 80% of CNVs we collapsed together shared more than 30% of their probes while 50% of them shared more than 70% of their probes (i.e.: probes identically assigned to a CNV as deletion or duplication by different segmentation algorithms), as shown on Figure 2.

Comparing algorithms, only 7% of the total burden of CNVs was common at all three algorithms while 18% of CNVs were found in at least two algorithms, corresponding to 35%, 74% and 34% of the respective totals of GTC, PENNCNV and QUANTISNP (see Table 5 and Figure 3). We developed our further analysis on these consensual CNVs.

Table 5. Percentage of CNVs common between different algorithms (percentages are referred to the total of the segmentation algorithm reported in column).

	GTC	PENNCNV	QUANTISNP
GTC	100%	68%	30%
PENNCNV	28%	100%	24%
QUANTISNP	21%	39%	100%
Consensual CNVs	35%	74%	34%
Total number of CNVs	9490	3934	6507

STRATIFICATION

Spurious association may be obtained or true association overlooked if allele frequencies differ notably between subpopulations that are represented unequally between cases and controls due to systematic ancestry differences. Population structure needs to be assessed to make reliable conclusions in association studies.[77-79] Most study designs propose sampling cases and controls from groups that share the same geographic origin or self-reported ethnic background, with the implicit assumption that no substructure exists within such groups. Results, however, indicate that analytic strategies need to take account of substructure as self-reported ethnicity is often inaccurate. Although we performed our initial analysis in reputed homogeneous population, we assessed for potential differences between cases and controls by measuring allelic frequencies of relatively common ethnic-specific alleles using the Ancestry Informative Markers (AIMs) set and by principal component (PCA) analysis. Specifically, we used the PCA method by Price et al. as implemented in software package EIGENSTRAT to control for population stratification.[80,81] This method enables detection and correction of population stratification on a genome-wide scale by using PCA to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. This simple, efficient approach performs well under various situations including both discrete mixed populations and admixed populations and is more powerful than the genomic control approach and the STRUCTURE approach.[82,83]

Our analysis resulted in ten significant principal components as calculated with twstats, however most of the variance was captured by the first four principal components (as upon the inflexion of eigenvalues scree plot). Correcting with these four will lead to acceptable correction in further analysis as shows the Q-Q plot on Figure 4.

IMPUTATION

Imputation of untyped SNPs was done using Mach v. 1.0.16 and a reference panel composed of CEPH and TSI samples from the Hapmap consortium.[70]

After quality controls we carried out a two-step imputation process using the greedy option and 100 rounds of iterations. In the first step a model is build that relates samples to the haplotypes in the reference panel and includes both an estimate of the "error" rate for each marker (an omnibus parameter which captures both genotyping error, discrepancies between typing platform and the reference panel, and recurrent mutation) and of "crossover" rates for each interval (a parameter that describes breakpoints in haplotype stretches shared between samples and the reference panel). In the

second step the parameters estimated in the first step are used to impute all SNPs in the reference panel in the sampled individuals.

A third step of quality control consisted in removing all imputed SNPs that had a squared correlation with true genotypes below 0.30, to eliminate most of the poorly imputed SNPs but only a small number of well imputed ones.

From 647,632 SNPs, we reached 1,413,093 SNPs after imputation, for a total genotyping rate of 0.957 in 345 individuals. We could finally analyze a total of 1,253,352 SNPs (after filtering for squared correlation and MAF).

For data management, we used the convert_mach.pl script from the GENGEN suite made available from the Open Bioinformatics software repository to reformat MACH imputed data.[69,84] We included it in an extensive UNIX syntax script that performed overall imputation process, including strand flipping and verification through LD structure with the --flip-scan option in PLINK.[73] The former process calculates the signed correlation between each SNP and a set of nearby SNPs in the reference and studied dataset separately to identify pairs of SNPs in which the absolute value of the genotypic correlation is above some threshold, so that counting the number of times the signed correlation is different in sign between the two datasets (a negative LD pair) versus the same (a positive LD pair) allows to finally eliminate outliers for such index.

CORRECTION

CNVs

We applied regional GC correction to correct for the technical genome-wide artifact known as GC wave or waviness, which is a common, systematic issue observed with whole-genome assays. It corresponds to a spatial wave pattern in log2 ratios, exhibiting larger than expected variations of log2 ratios, that appears to be genome-wide or chromosome specific in some samples and is generally highly correlated with the GC content of the genomic region surrounding the probes.[85,86] Data from literature and reported cases indicate that regions within the wave have an increased number of segments of copy number change and that these segments are longer than typical CN segments found in normal populations.[86] Briefly, to apply GC correction, for each sample markers are divided into 25 bins based on the equally spaced percentiles of the average GC count (GC content)from 250 kb upstreamto250 kb downstream of a particular marker. For the autosomal markers in each bin, the median log2 ratio of each bin is adjusted to zero and interquartile ranges (IQRs) are equalized across all the bins. The IQRs of all the adjusted log2 ratios (including the X and Y chromosomes) are multiplied by a factor that makes the IQRs of the adjusted log2 ratios equal to the IQRs of the original log2 ratios. Finally, the log2 ratios of all markers (including X and Y markers) in a bin are adjusted using the median of the medians of the log2 ratios for each autosome.

SNPs

For subsequent analysis we considered the following correctors in our model: the first 4 principal components, and imputation probability.

Principal components were derived from the smartpca.perl and twstats scripts from the EIGENSTRAT v. 3.0 package, to respectively calculate principal components and determine their statistical significance, as described before. Imputation probabilities were derived from MACH v. 1.0.16 outputs, as described before.[69,81]

ASSOCIATION TESTING

CNVs

To consider only rare CNVs, all CNVs present in more than 1% of TD samples were removed from association testing analysis. Furthermore, testing the one-sided hypothesis of an increase of rare CNVs in NTD samples, we removed all CNVs non present in NTD samples.

Case-control analysis for association with TD status was done using permutation procedures made available in PLINK v. 1.07.[73] Permutation procedures have desirable properties such as relaxing assumptions about normality and providing a framework for correction for multiple testing, which drove our choice.

Results of the analysis were visually checked using WGAVIEWER and compared against known copy number polymorphisms to further select interesting CNVs (containing known genes or small RNA sequences).[87]

Association of rare CNVs with TD status was tested using 100,000 permutations, on the basis of markers as well as gene based (as upon the --cnv-test-regions commands in PLINK).[73] The marker based analysis resulted in one duplicated region nearly significant at genome-wide level, while the gene based approach gave similar results (i.e.: first ranking genes were in the same regions as in the marker based analysis) but not genome-wide significant: genome-wide corrected p=0.075 (gw-p) with point-wise p=0.012 (pw-p) upon marker based analysis, and gw-p=0.469 with a pw-p=0.014 upon gene based analysis. This region, situated on chromosome 4 and spanning from 165.4 to 166.5 Mbp crosses the TRIM60, TRIM61 and KLHL2 genes. The signal in this region showed a coherent and well-shaped duplicated signal in two NTD subjects, not present in TD subjects as shown on Figure 5, and was found by all three segmentation algorithms, with one segmentation algorithm even finding a scarce signal for an additional NTD patient as shown on Figure 6.

SNPs

Time to event was calculated as the time between birth and the first blood transfusion for each patient, if this occurred, for a lifelong period until last follow-up (January 2010). Predictive factors considered were gender, SNP genotype (coded as 0, 1 or 2 upon the number of copies of the minor allele, while 0 and 1 for SNPs on X chromosomes for males) and imputation probability.

Considering the longitudinal nature of the study, independent variables were entered into a Cox proportional hazards model to identify significant predictive factors for each outcome. Patients were considered uncensored when blood transfusion happened during the study (TD patients) and censored when blood transfusion did not happen (NTD patients).

To do so we used the coxph() function from the R package SURVIVAL.[71,88] In its basic form the coxph() model attempts to fit survival data with covariates z to a hazard function of the form:

$$h(t|z) = h_0(t)e^{\beta' z}$$

where β is an unknown vector and $h_0(t)$ is the non-parametric baseline hazard. To find the parameter β, the partial likelihood is solved in the form:

$$L(\beta) = \prod_{i=1}^{D} \frac{e^{\beta' z_{(i)}}}{\sum_{j \in R(t_i)} e^{\beta' z_j}}$$

where $R(t_i)$is the risk set at time t_i.

Given the estimate of β, $\hat{\beta}$ (a vector), along with the estimates' covariance matrix, $\hat{I}^{-1}$, $\hat{\beta} \sim AN(\beta, \hat{I}^{-1})$ holds approximately since $\hat{I} \to I$ as $n \to \infty$ and this approximation allows local tests to be done. A local null hypothesis can usually be put into matrix form, $C\beta = d$ where C is a $q \times p$ matrix of full rank and d is a vector of length q. Under this setup, the test statistic is:

$$X_w^2 = (C\hat{\beta} - d)' \left[C\hat{I}^{-1}C'\right]^{-1} (C\hat{\beta} - d)$$

which under the null hypothesis follows χ_q^2 (this is the Wald test).

Accordingly, the coxph() function includes in its summary:

- estimates of the β_k including standard errors and p-values for each test $H0 : \beta_k = 0$ with the other $\beta_j = \hat{\beta}_j$,
- estimate of the risk ratio with confidence bounds,
- p-values for likelihood ratio, Wald and score tests for the global null, $H0 : \beta_i = 0$ for all i.

We applied the coxph() function to our data through a R plugin for PLINK, and considered statistically significant all SNPs with p-values less than 1.00E-07 (corresponding to a genome-wide significance level of 0.05 after a Bonferroni correction for multiple testing of 500,000 independent SNPs).[71,73]

The SNPs were then plotted in their genomic context using WGAVIEWER software, ordered upon their p-values and annotated against the last versions of genomic maps using ENSEMBL database (homo sapiens release 60, GRCh37).[87,89] Finally a comprehensive annotation of surrounding SNPs with their respective LD measure was done, to assert the coherence of statistically significant signals. Figure 7 gives an overview of results per chromosome.

GENOME-WIDE SIGNIFICANT RESULTS

Among genome-wide significant results (summarized in Table 6), the first one is BCL11A (most significant SNP: rs766432, p=2.24E-09), a well-known gene responsible of HbF production increase. Such result confirm the coherence of our approach and the effect of BCL11A not only on HbF production but as a major determinant of homozygous β°-thalassemia phenotype. The second ranking SNP, rs2541639, within HBM gene in the α-globin gene cluster is actually partially correlated to α gene deletions responsible for the observed phenotypic variation (see after for a detailed analysis of α-globin gene cluster). The other gene significant at genome-wide level (PTPRD - rs7861848 with p=1,66E-08) is not known to influence HbF levels or β-thalassemia phenotype but encodes a protein tyrosine phosphatase, and members of this family are known to regulate a variety of cellular processes including cell-growth, differentiation, mitotic cycle and oncogenic transformation. However the SNP presents a low minor allele frequency (1%) and is isolated as all correlated SNPs were excluded for presenting even lower minor allele frequencies. This result is therefore not reliable.

Table 6. Genome-wide significant results.

SNP	Rank	p-value	Chr.	Position	Gene
rs766432	1	2.20E-09	2	60,719,970	BCL11A
rs2541639	2	1.00E-08	16	205,035	HBM
rs1427407	3	1.48E-08	2	60,718,043	BCL11A
rs7861848	4	1.66E-08	9	8,753,534	PTPRD
rs10172646	5	1.79E-08	2	60,720,757	BCL11A
rs11886868	6	1.81E-08	2	60,720,246	BCL11A
rs10195871	7	1.81E-08	2	60,720,589	BCL11A
rs6732518	8	2.63E-08	2	60,708,597	BCL11A

RESULTS UNTIL HBS1L-MYB INTERGENIC REGION MOST SIGNIFICANT SNP (P=2.80E-05)

Beyond BCL11A and α-globin gene cluster present in genome-wide significant results, selecting SNPs with sufficient heterozigosity that demonstrate genome-wide or suggestive significance (until the p-value of the most significant SNP in the HBS1L-MYB intergenic region, already known for its influence on HbF – rs9399137, p=2.80E-05), some further genes can be described (see Table 7).

Table 7. Results with suggestive significance.

SNP	p-value	Chr.	position	MAF	Gene	
rs766432	2.20E-09	2	60,719,970	0.17	BCL11A	known HbF levels modifier
rs2541639	1.00E-08	16	205,035	0.39	HBM	α-globin gene cluster
rs17462275	3.22E-07	14	33,609,861	0.22	NPAS3	Neuronal PAS domain protein 3
rs10266	3.76E-07	16	107,162	0.4	SNRNP25 (3'UTR)	α-globin gene cluster
rs17525396	6.43E-07	16	145,884	0,20	C160rf35 / NPRL3	α-globin gene cluster
rs1288829	1.56E-06	3	71,334,477	0,46	FOXP1	Forkhead box transcription factor family
rs2562147	1.82E-06	16	114,535	0.48	RHBDF1	α-globin gene cluster
rs9838898	7.54E-06	3	152,905,655	0.43	RAP2B (downstream) (rs6785014 in 3'UTR: p=3.08E-05)	Ras family small GTP binding protein
rs1211375	1.23E-05	16	240,280	0.47	LUC7L	α-globin gene cluster
rs11958405	1.39E-05	5	22,247,159	0.49	CDH12	Cadherin 12, type 2 (N-cadherin 2)
rs11121928	1.54E-05	1	12,615,888	0.27	DHRS3 (downstream)	Dehydrogenase/reductase (SDR family) member 3
rs1899398	1.71E-05	3	3,770,669	0.15	SUMF1	Sulfatase modifying factor 1
rs6843966	1.87E-05	4	6,019,743	0.47	JAKMIP1 (downstream)	Janus kinase and microtubule interacting protein 1
rs675046	1.89E-05	11	107,176,219	0.37	CWF19L2 (downstream)	CWF19-like protein 2, cell-cycle control (S. pombe)
rs3788504	2.23E-05	22	33,160,397	0.07	SYN3	Synapsin III
rs7763010	2.37E-05	6	117,762,218	0.13	GOPC	Golgi associated PDZ and coiled-coil motif containing
rs1362512	2.51E-05	20	15,651,887	0.38	MACROD2	MACRO domain containing 2
rs9399137	2.80E-05	6	135,419,018	0.19	HBS1L-MYB intergenic region	known HbF levels modifier

Among these BCL11A, FOXP1, RAP2B, CDH12, SUMF1 and HBS1L-MYB intergenic region are the only loci showing coherent and at least partially independent signals, with their most significant marker mapping within the gene. Genomic context of these most relevant signals and their effect on phenotype severity are described in Figure 8, Figure 9, Figure 10,Figure 11, Figure 12, and Figure 13 respectively.

α-GLOBIN GENE CLUSTER

Among the results with genome-wide and suggestive significance we find many SNPs within the α-globin gene cluster, as rs2541639 the most significant one (p=1.00E-08) and rs10266 a non synonymous SNP within SNRNP25 (p=3.76E-07), together with (until p=5.0E-04): rs17525396 (p=6.43E-07), rs2562147 (p=1.28E-06), rs216605 (p=2.71E-06), rs1211375 (p=1.23E-05), rs2157115 (p=1.00E-04), rs2974771 (p=1.00E-04), rs6600233 (p=2.00E-04) and rs2541675 (p=2.00E-04). Interestingly these SNPs highly correlate with α mutation genotype, therefore loosing significance when accounting for such mutations in a model, except for rs6600233 (within C16orf35) that keep an independent effect on time to first transfusion (p=0.007, exp(β)=1.33) when accounting for top SNPs from BCL11A and HS1L-MYB region as well as α mutation genotype and gender, in a Cox proportional hazards model. To consider the effect of BCL11A and HBS1L-MYB intergenic region we used the following SNPs based on the work by Galarneau et al. (and conducing stepwise conditional analysis as well):

- for BCL11A: rs766432 (that is in full LD with rs4671393), rs10189857, and rs10195871 (the best tag for rs7599488);
- for HBS1L-MYB intergenic region: rs9399137 (even if we had rs4895441 with r^2=0.95 with rs9402686, as rs9399137 showed a stronger effect in our sample even if we have a lower r^2 between rs9399137 and rs9402686: r^2=0.86), rs11759077 (the closest to rs28384513 showing independent effect through conditional analysis), and rs9376092 and rs6920211 (the closest SNPs we had available around ss244317976).[25]

Therefore, thanks to the independent effect of rs6600233 we can hypothesize the presence in our sample of elements with a significant role in α-globin gene expression, independent from α-globin gene deletions or point mutations.

Genomic context of α-globin gene cluster and effect of α gene deletions are given in Figure 14, while some details of outcomes for rs6600233 are given in Figure 15: survival functions for each genotype while accounting for gender, BCL11A effect (only rs766432 and rs10189857 remained at last step of the Cox proportional hazards model), HBS1L-MYB intergenic region effect and α gene deletions effect, showing a significant difference between curves (p=0.007, exp(β)=1.33).

GENE BASED JOINT TESTING

To further analyze these results we developed a gene based joint testing approach to identify gene associated to the extreme form of the phenotype, using TD versus NTD status. We used a permutation based approach as made available in PLINK to identify genes with more than two independent SNPs (below r^2=0.5 calculated on our sample) at least nominally significant.[68] None of these passed a multiple testing Bonferroni correction but the first gene was BCL11A, while it is interesting to note that the gene FRMPD4, located within Xp22.3, ranked fourth with such approach (while ranking first in a case/control TD versus NTD association study, with rs5978532 mapping within the first intron of FRMPD4 gene – data not shown).

CANDIDATES FROM LITERATURE AND FUNCTIONALLY RELEVANT SNPs

To complete our analysis of results, we further report the p-values of known candidates from literature and functionally relevant SNPs until p=5.00E-04.

To retrieve known candidates from literature we used PHENOPEDIA and GWAS CATALOG databases as well as bibliographic resources from PUBMED, to cross them with our results.[90-92] From this analysis we can list:

- BCL11A with rs766432 as the most significant SNP, as previously mentioned (see before and Figure 8 for details).
- HBS1L-MYB intergenic region with rs4895441 as the most significant SNP, as previously mentioned (see before and Figure 13 for details).
- TSHZ2 with rs200603 (p=5.00E-04) at position 51,952,501bp on chromosome 20. This gene, that might encode for a transcriptional regulator involved in developmental processes, has been reported to be associated with Hb levels in a meta-analysis on erythrocyte phenotypes.[93]
- rs17525396 (within C16orf35, in the α-globin gene cluster and containing the α major regulatory element) has already been described to be associated to HbA2 and MCH phenotype in a Sardinian population in which, however, α gene deletions were not accounted for, while this SNP results highly correlated to α gene deletions in our sample.[59]
- rs9820070 (p=2.00E-04), an intergenic SNP at position 187,687,074 on chromosome 8. This SNP resulted associated with blood urea nitrogen in a recent GWAS on hematological traits in a Japanese population.[94]
- FLT1 (Fms-related tyrosine kinase 1) with rs12429309 (p=3.00E-04), a receptor for vascular endothelial growth factor (VEGF), has tyrosine protein kinase activity that is important for the control of cell proliferation and differentiation. This gene has been found to be the most associated with HbF percentage levels in a cohort of 137 patients with sickle cell anemia in which 280 SNPs were tested for association with HbF response to hydroxyurea treatment.[95]
- Regarding the 8q QTL described by Garner et al., we could not investigate such candidate as it is supposed to interact with the *Xmn*I-Gγ polymorphism, which is not represented in the present sample. For information the most significant signal in an area including 1Mbp before D8S538 and 1Mbp after D8S1833 is rs7843301 within SNTG1, a member of the syntrophin family specifically expressed in the brain.[51-53]
- Regarding the Xp22.3 QTL, no significant association with the time to first transfusion was found. However developing a case/control TD versus NTD association study on SNPs a signal in the FRMPD4 gene was found with rs5978532 ranking first (see to the previous gene based joint testing section also).[49]

Regarding SNPs described as functionally relevant, Table 8 list them until p=5.00E-04.

Table 8. Functionally relevant SNPs.

SNP	p-value	Chr.	Position	Type	Gene
rs10266	3,76E-07	16	107,162	RR - 3' UTR	SNRNP25
rs216605	2,71E-06	16	113,924	SS - Intronic	RHBDF1
rs17201603	1.00E-04	3	36,898,973	NSC	TRANK1
rs591157	2.00E-04	10	96,954,298	NSC	C10orf129
rs12364019	3.00E-04	11	5,730,343	NMD Transcript	TRIM22
rs3732413	4.00E-04	3	119,133,183	NSC	ARHGAP31
rs2301487	5.00E-04	16	547,297	NSC	RAB11FIP3

[RR = Regulatory Region, NSC = Non Synonymous Coding, SS = Splice Site, NMD = Nonsense Mediated Decay].

MODELLING OF PHENOTYPE AMELIORATION

To summarize and further investigate these results we decided to model the impact of such results on β°-thalassemia phenotype amelioration, developing a Cox proportional hazards model for the time to first transfusion. We included in the model: gender, BCL11A effect, HBS1L-MYB effect, α gene deletions status, and all suggestively significant results from the present study. All markers were treated as having an additive effect.

Regarding BCL11A and HBS1L-MYB effect, as mentioned before, we used the following SNPs based on the work by Galarneau et al. and stepwise conditional analysis:

- for BCL11A: rs766432, rs10189857, and rs10195871;
- for HBS1L-MYB intergenic region: rs9399137, rs11759077, rs9376092 and rs6920211.[25]

We elaborated two Cox proportional hazards models: the first included only known modifiers of phenotype severity while the second added to the significant results of the first one, the SNPs with suggestive significance. We conducted the analysis in a stepwise manner considering two terms and three terms interactions within BCL11A and HBS1L-MYB intergenic region.

We considered in the first model: gender, BCL11A, HBS1L-MYB intergenic region, α gene deletions status and rs6600233, the latter being the previously found independent signal within the α-globin gene cluster, a known modifier of β-thalassemia phenotype. Gender did not influence the outcome and no interactions were significant between independent SNPs within BCL11A and HBS1L-MYB intergenic region. Among SNPs introduced at first step rs10189857, rs10195871 (both within BCL11A), rs9376092 and rs6920211 (both within HBS1L-MYB intergenic region) did not show significant effect on the outcome, while all others kept independent effects on the time to first transfusion with the strongest effect for rs9399137, within HBS1L-MYB intergenic region. Results presented in Table 9 demonstrate that BCL11A, HBS1L-MYB intergenic region and α gene deletions status are the three main genetic determinants influencing severity of homozygous β°-thalassemia phenotype. Furthermore, another independent signal within α-globin gene cluster contribute to the phenotype severity.

The second model added to the first one the following SNPs: rs1899398 (SUMF1 - chr.3), rs1288829 (FOXP1 - chr.3), rs11958405 (CDH12 - chr.5), rs7763010 (GOPC - chr.6), rs17462275 (NPAS3 - chr.14), rs1362512 (MACROD2 - chr.20), rs3788504 (SYN3 - chr.22), rs11121928 (DHRS3 - chr.1), rs9838898 (RAP2B - chr.3), rs6843966 (JAKMIP1 - chr.4) and rs675046 (CWF19L2 - chr.11). All SNPs remained to final step of the model except rs1362512 and rs9838898. As we can see in Table 9, elements introduced in the first block did not change much when adding the suggestive results of the study. Without a replication of the suggestive results from the GWAS, the only interest of the second model is to establish the independency of effects of these SNPs. Figure 16 shows the survival function at the mean of covariates for model 2.

Table 9. Determinants of phenotype amelioration upon Cox proportional hazards model for time to first transfusion.

		β	Std. Err.	Wald stat.	df	p-value	exp(β)	exp(β) 95% IC	
								Inf.	Sup.
Model 1	rs766432 [BCL11A - chr.2]	-0.766	0.147	27.04	1	<0.0001	0.46	0.35	0.62
	rs9399137 [HBS1L-MYB - chr.6]	0.913	0.190	23.03	1	<0.0001	2.49	1.72	3.62
	α gene deletions [chr.16]	0.479	0.112	18.12	1	<0.0001	1.61	1.29	2.01
	rs6600233 [C16orf35 - chr.16]	0.267	0.105	6.41	1	0.0114	1.31	1.06	1.61
	rs11759077 [HBS1L-MYB - chr.6]	-0.414	0.169	6.01	1	0.0142	0.66	0.48	0.92
Model 2	rs9399137 [HBS1L-MYB - chr.6]	0.981	0.195	25.24	1	<0.0001	2.67	1.82	3.91
	rs7763010 [GOPC - chr.6]	0.671	0.151	19.75	1	<0.0001	1.96	1.45	2.63
	rs675046 [CWF19L2 - chr.11]	-0.417	0.099	17.59	1	<0.0001	0.66	0.54	0.80
	rs11121928 [DHRS3 - chr.1]	0.492	0.121	16.41	1	0.0001	1.64	1.29	2.08
	α gene deletions [chr.16]	0.462	0.118	15.20	1	0.0001	1.59	1.26	2.00
	rs766432 [BCL11A - chr.2]	-0.570	0.156	13.38	1	0.0003	0.57	0.42	0.77
	rs3788504 [SYN3 - chr.22]	0.723	0.204	12.57	1	0.0004	2.06	1.38	3.07
	rs11759077 [HBS1L-MYB - chr.6]	-0.579	0.176	10.83	1	0.0010	0.56	0.40	0.79
	rs11958405 [CDH12 - chr.5]	-0.311	0.100	9.71	1	0.0018	0.73	0.60	0.89
	rs17462275 [NPAS3 - chr.14]	-0.295	0.114	6.74	1	0.0094	0.74	0.60	0.93
	rs1288829 [FOXP1 - chr.3]	-0.253	0.106	5.65	1	0.0174	0.78	0.63	0.96
	rs6843966 [JAKMIP1 - chr.4]	-0.252	0.108	5.48	1	0.0193	0.78	0.63	0.96
	rs1899398 [SUMF1 - chr.3]	-0.291	0.144	4.07	1	0.0436	0.75	0.56	0.99
	rs6600233 [C16orf35 - chr.16]	0.211	0.113	3.50	1	0.0613	1.24	0.99	1.54

CONCLUSIONS

Without treatment, children with thalassemia major have severe failure to thrive and shortened life expectancy. Furthermore, even if treatment with a regular transfusion program and chelation therapy extends life expectancy into the third to fifth decade, these individuals are at risk for iron overload secondary to increased intestinal absorption of iron as a result of ineffective erythropoiesis. This excess is the cause of many serious complications affecting mostly liver, heart and endocrine glands, that can be fatal to patients.

Much can still be done to improve life-expectancy and quality of life of these patients and some of it probably lies in better understanding of phenotypic severity variations. A lot of work has been done in the recent years to explain HbF production variations, an important modulator of β-thalassemia phenotype severity. However, a great part of phenotypic variation has still to be clearly understood and its eventual genetic determinants to be unrevealed.

To our knowledge, this study is the first whole genome association scan for modifiers of homozygous β°-thalassemia phenotype severity (through survival analysis of time to first transfusion), as well as the first association study of rare CNVs with HbF levels (through NTD patients characterized by the presence of nearly 100% of HbF).

The results from the present study demonstrate that BCL11A, HBS1L-MYB intergenic region and α-globin gene deletions are major determinants of homozygous β°-thalassemia phenotype severity, using a genome-wide approach and an accurate phenotype definition on a particularly homogeneous population, therefore bringing detailed and new information to the previous works dealing with genetic determinants of β-thalassemia phenotype severity.[61,96,97] These results are coherent with findings on genetic determinants of disease severity in β°-thalassemia/hemoglobin E Thai patients from Nuinoon et al. as this study found only BCL11A and HBS1L-MYB, outside of the β-globin gene cluster, to be significantly associated with disease severity score.[61] Regarding the other GWAS conducted by Sherva et al. on the same population and phenotype (severity score was defined differently however), only the HBS1L-MYB intergenic region was found in common with the present study.[96]

Furthermore, we show that the severity of β°-thalassemia phenotype might be influenced by a rare CNV showing duplication on chromosome 4 spanning from 165.4 to 166.5 Mbp that was found in two NTD patients and no TD patient. The two patients presenting such CNV are negative for the BCL11A protective allele (according to rs766432) while one is positive for the HBS1L-MYB protective allele (according to rs9399137), leaving unexplained at least for one patient his NTD status (we do not consider α-globin gene deletions as the case/control study applied for CNV association investigates HbF levels). The CNV crosses three genes, namely TRIM60, TRIM61 and KLHL2. The two first ones encode for proteins that contain a RING finger domain, a motif present in a variety of functionally distinct proteins and known to be involved in protein-protein and protein-DNA interactions. KLHL2 is a member of the kelch family of protein, which members associate with the actin cytoskeleton and regulate process length, but KLHL2 is also suspected to be a tumor growth promoter. Furthermore, in addition to the genes crossed, the microRNA MIR578 results to be produced in this region. MIR578 has BCL11B among its targets, a gene which function is still unknown but showing a high homology with BCL11A. It could therefore present the same sequence as the one in BCL11A gene, responsible for the formation of a repressor complex.[43]

We also show that a SNP within the α-globin gene cluster determine an effect on phenotype independent from α-globin gene deletions or point mutations. Thanks to this SNP we can hypothesize the presence in our sample of elements with a significant role in α-globin gene expression, independent from α-globin gene deletions or point mutations. This could involve deletions of the α major regulatory element (α-MRE) as it lies 10kbp downstream from rs6600233, or even a mechanism

similar to the creation of a new promoter-like as described by De Gobbi et al.[98,99] Such result is sustained by recent studies demonstrating that α-MRE polymorphisms are correlated with variations in gene expression, although the SNPs constituting the six known different haplotypes are outside the binding sites for nuclear factors.[100]

Finally, the present study also gives some elements about twelve other markers that could be involved in homozygous β°-thalassemia phenotype severity, of which four present a reliable signal within a gene and three also show an independent effect when modeled with other determinants of β-thalassemia phenotype severity (the four genes being FOXP1, CDH12, SUMF1 and RAP2B). These genes are not significant at genome-wide level but their function is relevant to β-thalassemia phenotype severity.

FOXP1 gene (forkhead box P1) belongs to the subfamily P of the forkhead box transcription factor family. Forkhead box transcription factors play important roles in the regulation of tissue and cell type-specific gene transcription during both development and adulthood. This ubiquitously expressed gene is implicated in malignancies and plays an important role in cardiac and lung development, B-cell development and macrophage differentiation.

CDH12 gene (cadherin 12, type 2 - N-cadherin 2) encodes a type II classical cadherin from the cadherin superfamily of integral membrane proteins that mediate calcium-dependent cell-cell adhesion. This particular cadherin appears to be expressed specifically in the brain but its mRNA is present in several additional tissues, suggesting that there is posttranscriptional control of this gene's expression.[101] Mutations and loss of expression of cadherins have been implicated in the progression of some malignant tumors, suggesting that cadherins including CDH12 may act as tumor-metastasis suppressor genes.[102,103]

SUMF1 gene (sulfatase modifying factor 1) encodes an enzyme that catalyzes the hydrolysis of sulfate esters by oxidizing a cysteine residue in the substrate sulfatase to an active site 3-oxoalanine residue, which is also known as C-alpha-formylglycine. Mutations in this gene determine cellular dysfunction by blocking autophagic protein clearance.[104] SUMF1 is therefore correlated to the unfolded protein response (UPR) as the autophagic process is required to counterbalance the expansion of endoplasmic reticulum during the UPR.[105-108] And prolonged UPR activation induces apoptosis, so that SUMF1 could be correlated to erythroptosis.[109] Furthermore SUMF1 seems to control hematopoietic stem progenitor cells differentiation and hematopoietic lineage development through FGF and Wnt signaling.[110]

RAP2B encodes for a Ras family small GTP binding protein expressed in human red blood cells where it might have a role in membrane shedding, while the loss of phosphatidylserine asymmetry of the erythrocyte membrane which plays key role in early clearance of erythrocytes from circulation in β°-thalassemia, seems to be mediated by membrane vesiculation.[111,112] Furthermore RAP2B is also believed to interact with platelet cytoskeleton by direct binding to actin filament, while the existence of a chronic hypercoagulable state involving significant changes in platelet membrane protein fractions is recognized in β thalassemic patients.[113,114] Changes in protein fractions of both erythrocytes and platelets correlate with severity of clinical manifestation of the disease.[115]

These results need to be further investigated, eventually through sequencing of the α-MRE, laboratory validation of the CNV, and replication of the suggestively significant SNPs in an independent cohort of β thalassemic patients. Hopefully the duplication on chromosome 4 could bring new insights on HbF production mechanisms while some of the most significant SNPs might uncover aspects of the unexplained phenotypic variability of β-thalassemia, and make a step towards better life for patients.

FIGURES

Figure 1. Identity by state of samples.

[back to text]

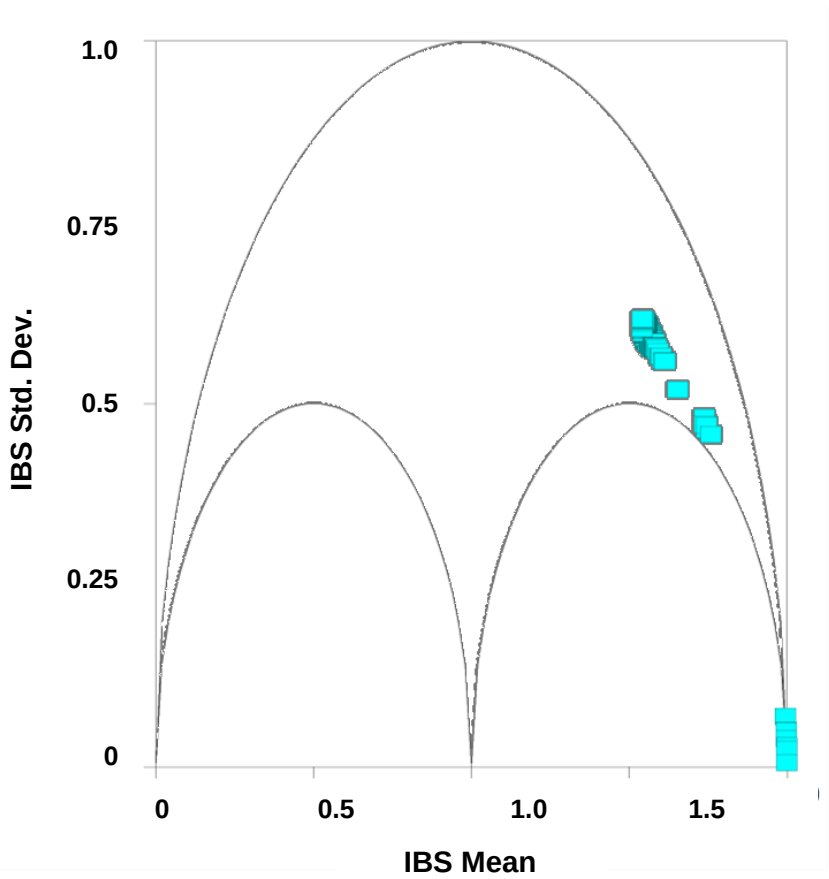


Figure 2. Percentage of overlap between CNVs collapsed together when identifying consensual CNVs between segmentation algorithms.

[back to text]

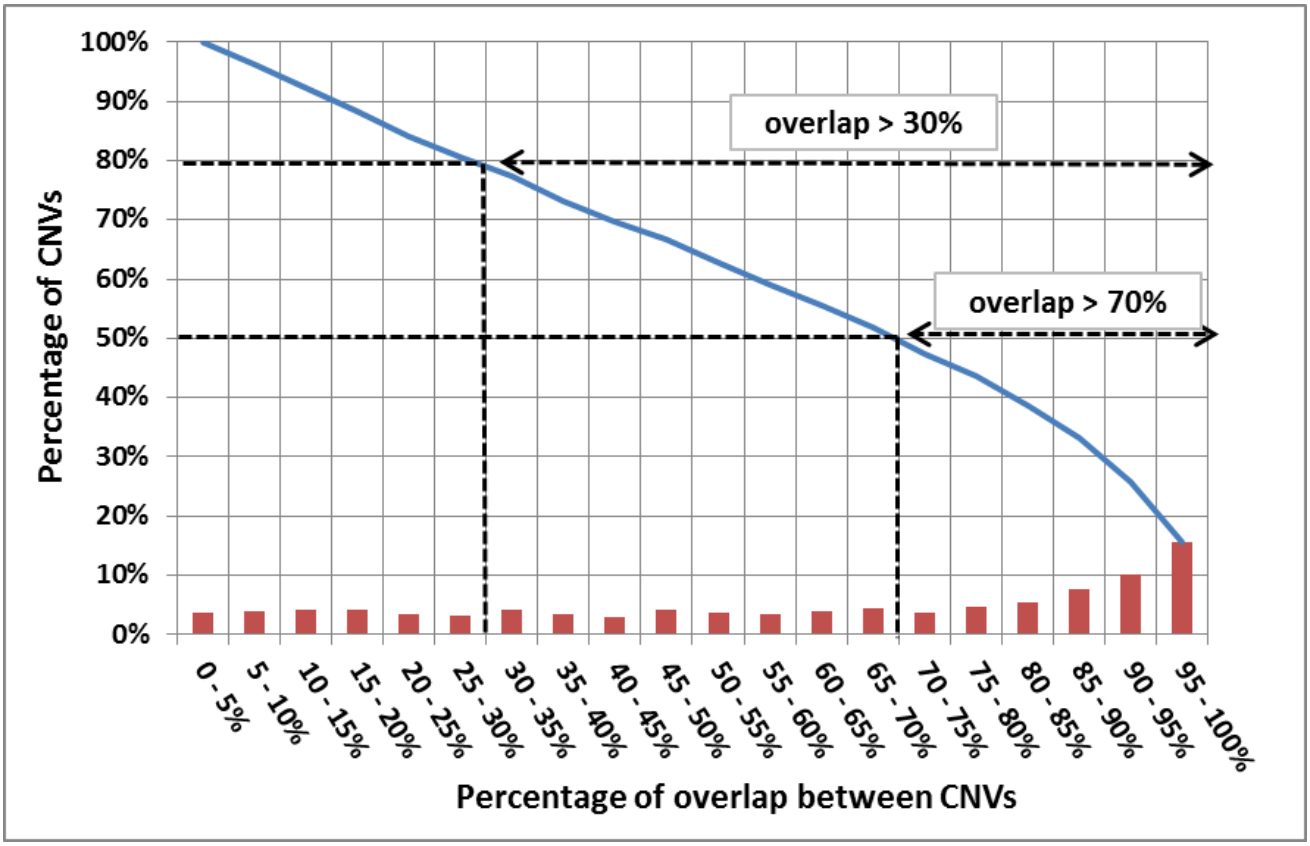


Figure 3. Consensual CNVs between segmentation algorithms.

[back to text]

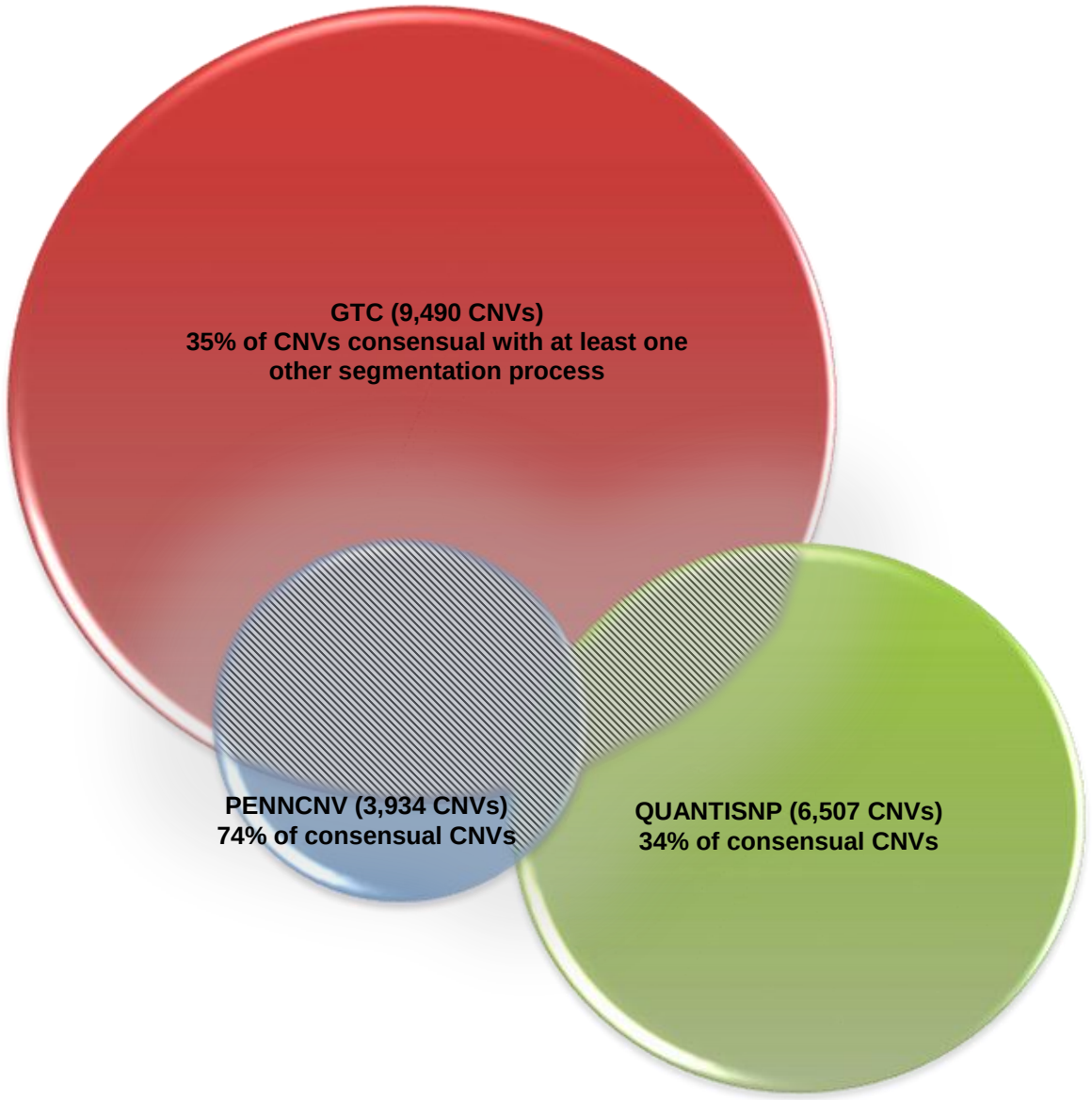


Figure 4. Quantile-quantile plot of observed associations compared with the expectations under no association.

[back to text]

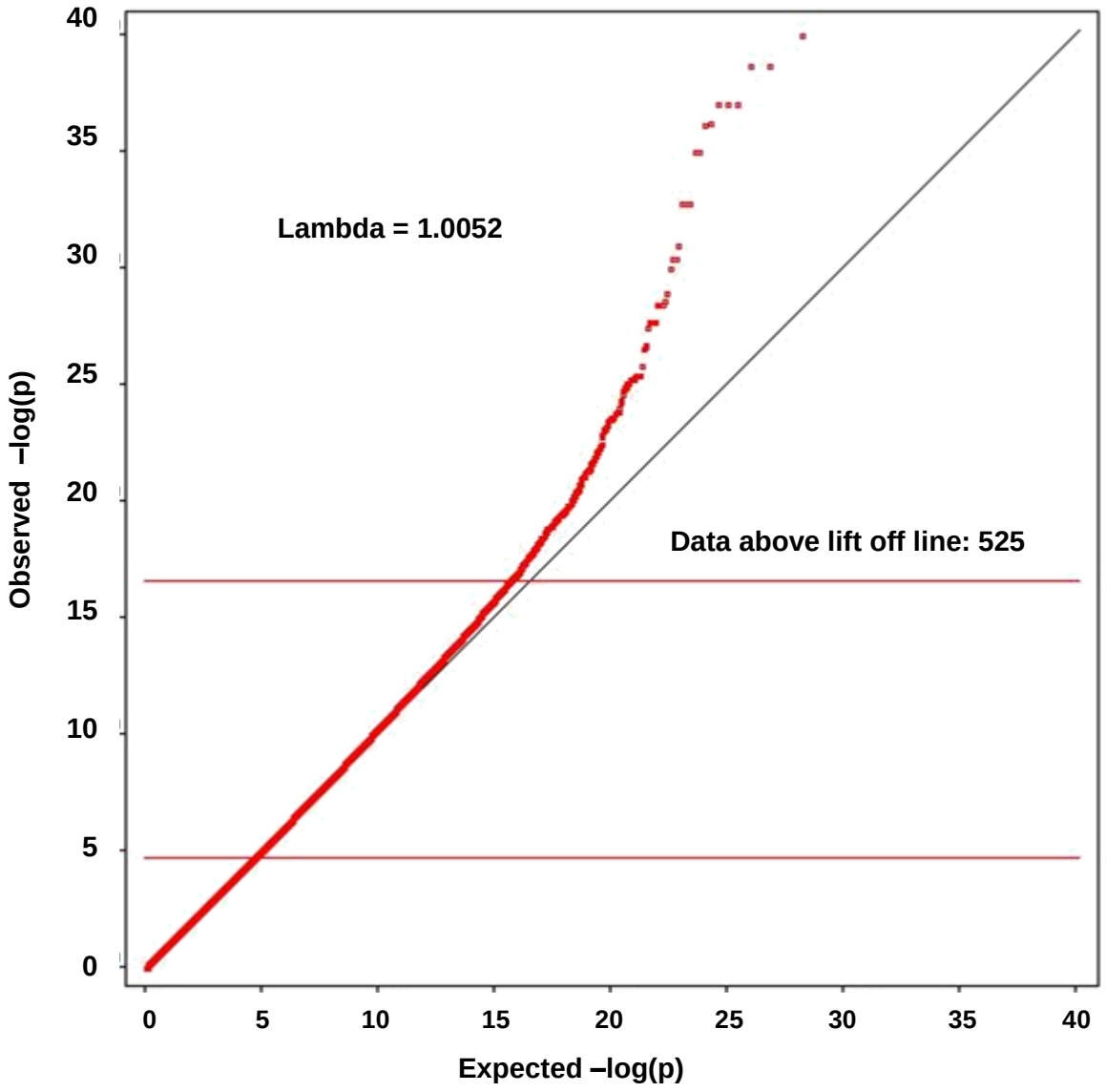


Figure 5. Intensity signals of two NTD samples showing duplication in a region spanning from 165.4 to 166.5Mb on chromosome 4.

[back to text]

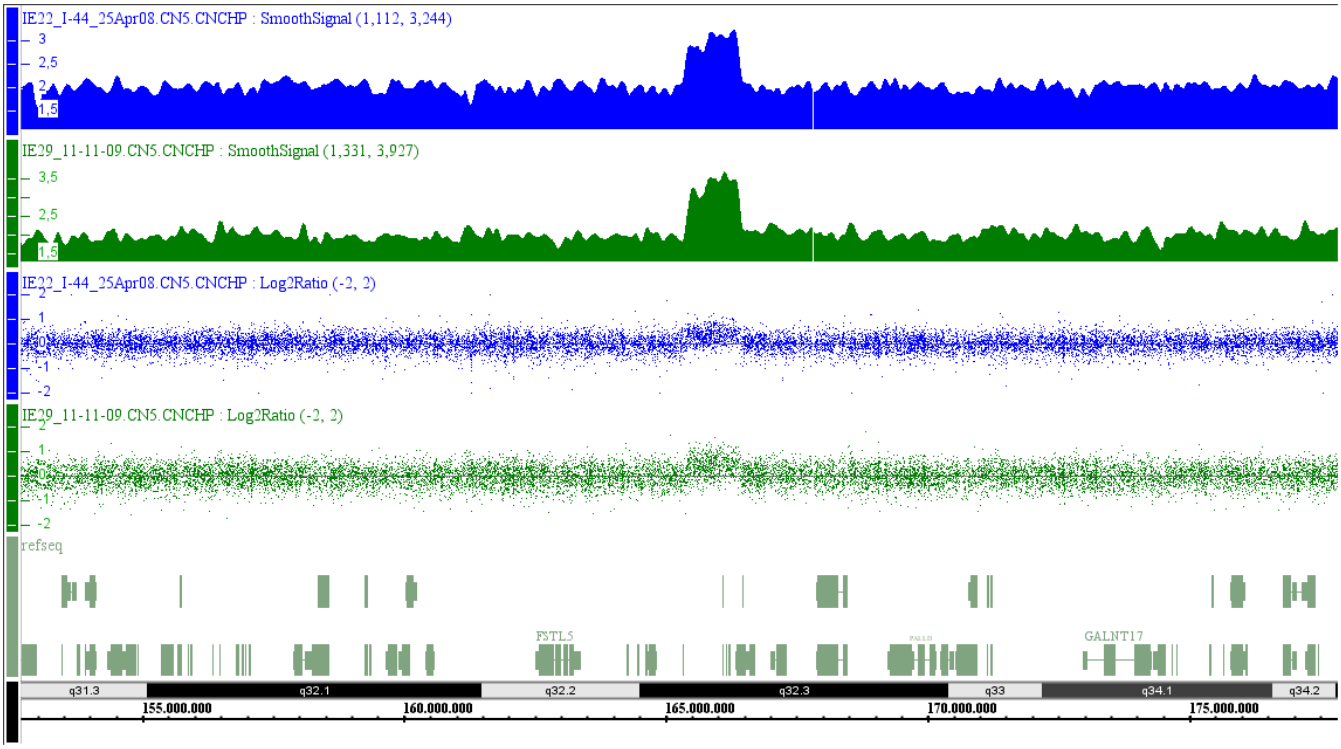


Figure 6. Intensity signals of NTD samples showing duplication in a region spanning from 165.4 to 166.5Mb on chromosome 4, upon different segmentation algorithms.

[back to text]

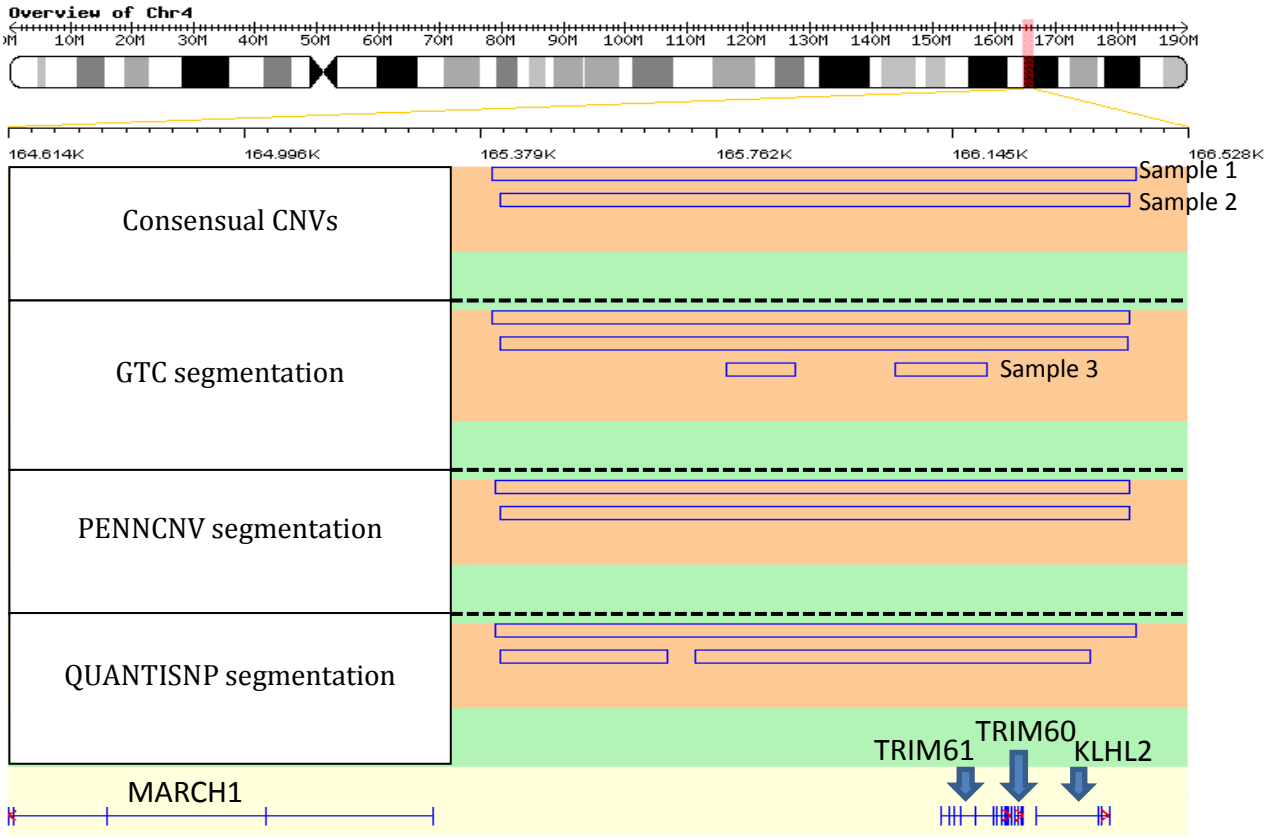


Figure 7. Manhattan plot.

[back to text]

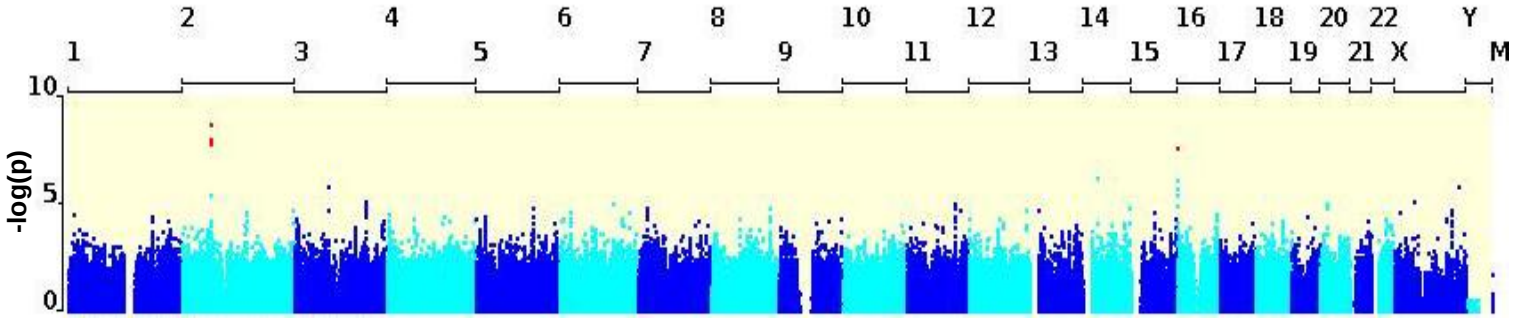


Figure 8. BCL11A gene and effect of its respective most significant SNP on time to first transfusion.

[back to text]

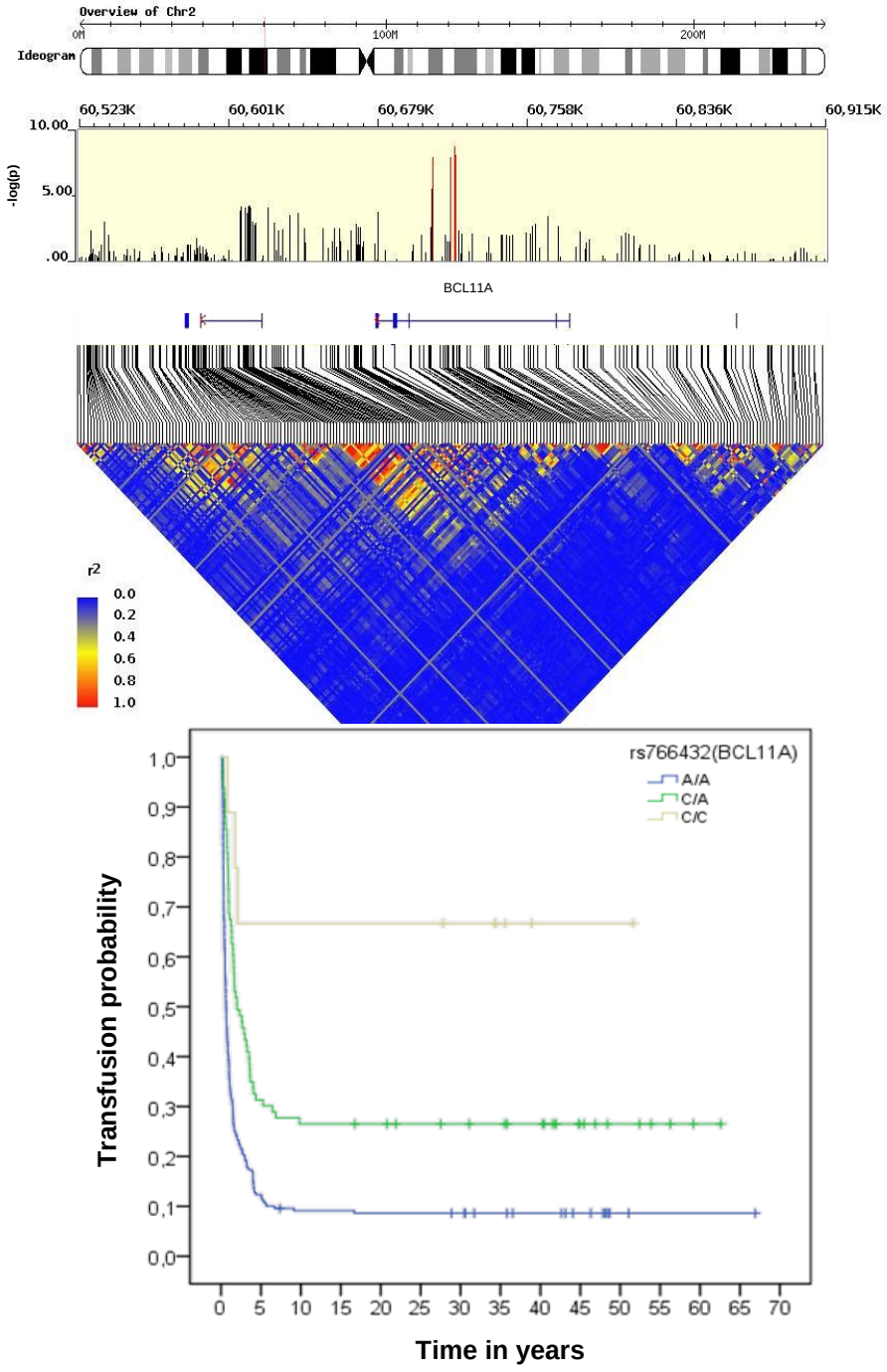


Figure 9. FOXP1 gene and effect of its respective most significant SNP on time to first transfusion.

[back to text]

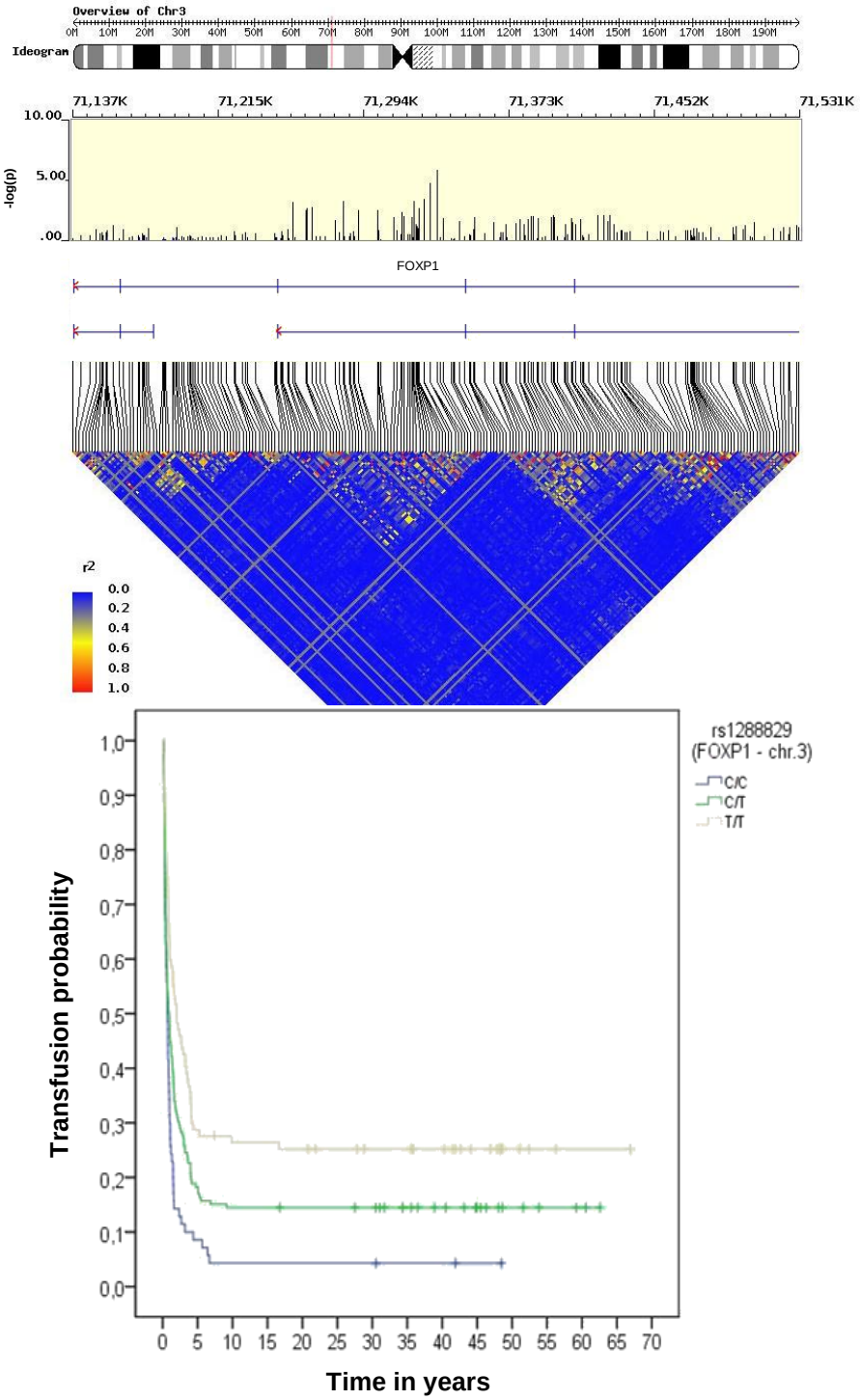


Figure 10. RAP2B gene and effect of its respective most significant SNP on time to first transfusion.

[back to text]

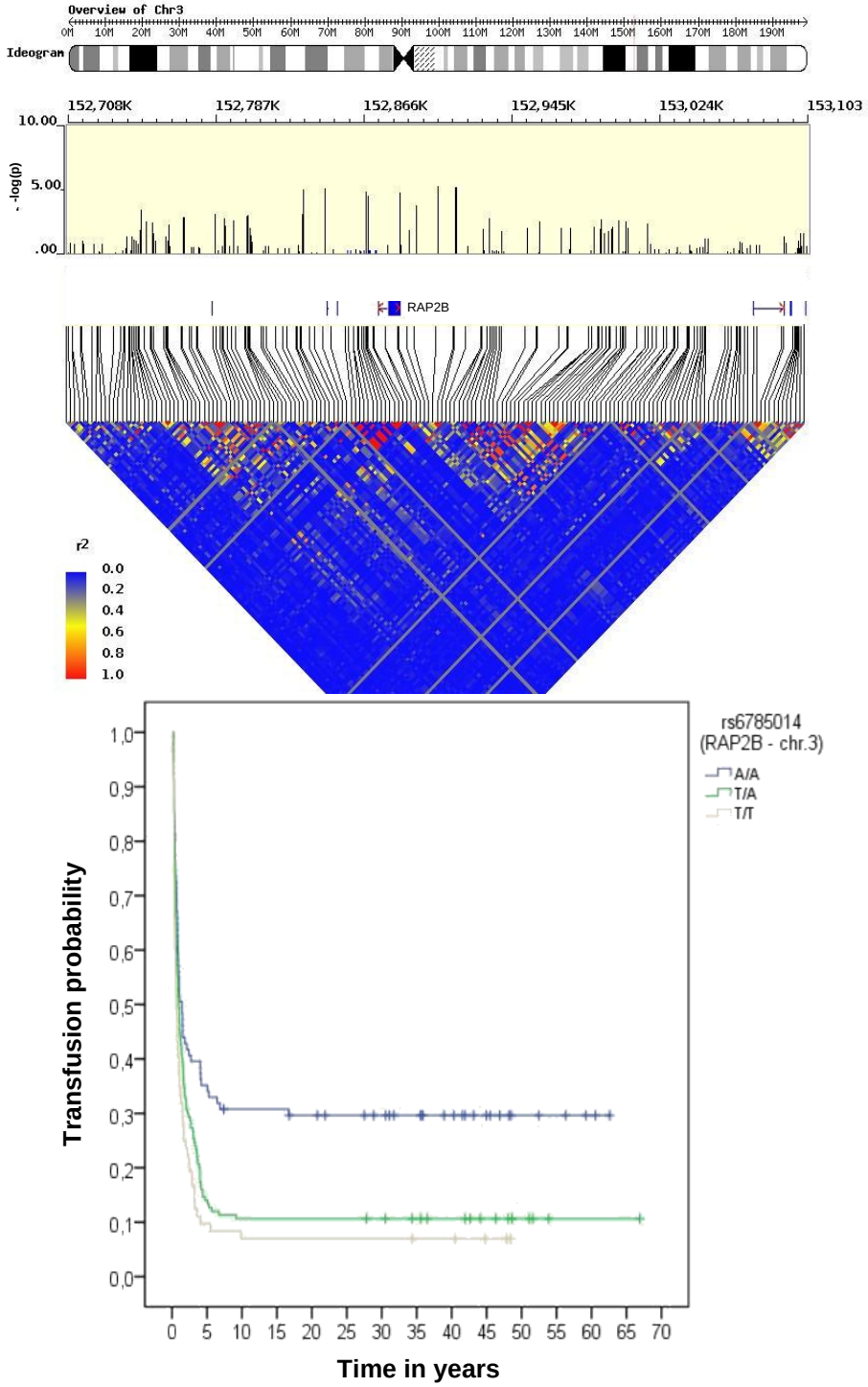


Figure 11. CDH12 gene and effect of its respective most significant SNP on time to first transfusion.

[back to text]

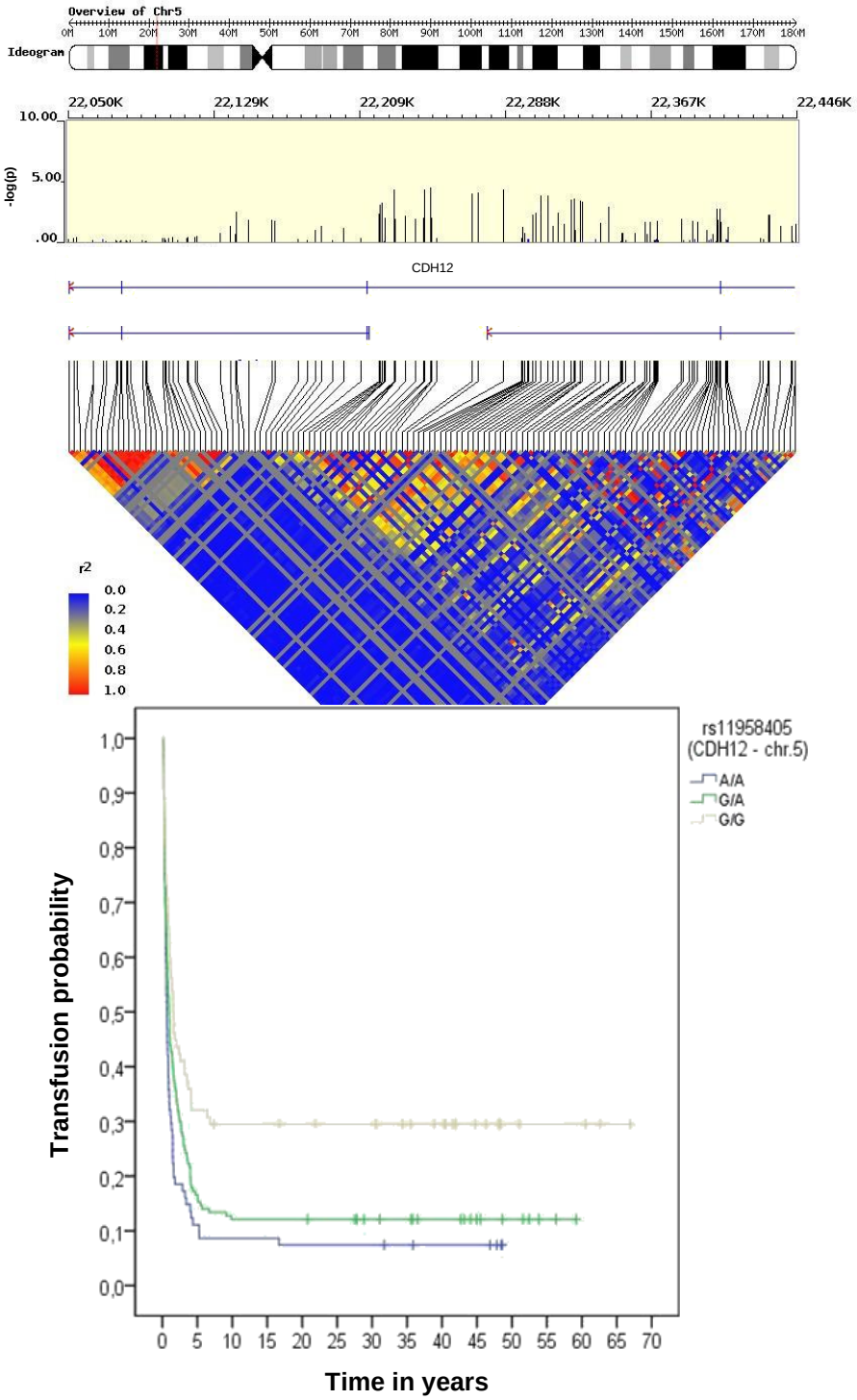


Figure 12. SUMF1 gene and effect of its respective most significant SNP on time to first transfusion.

[back to text]

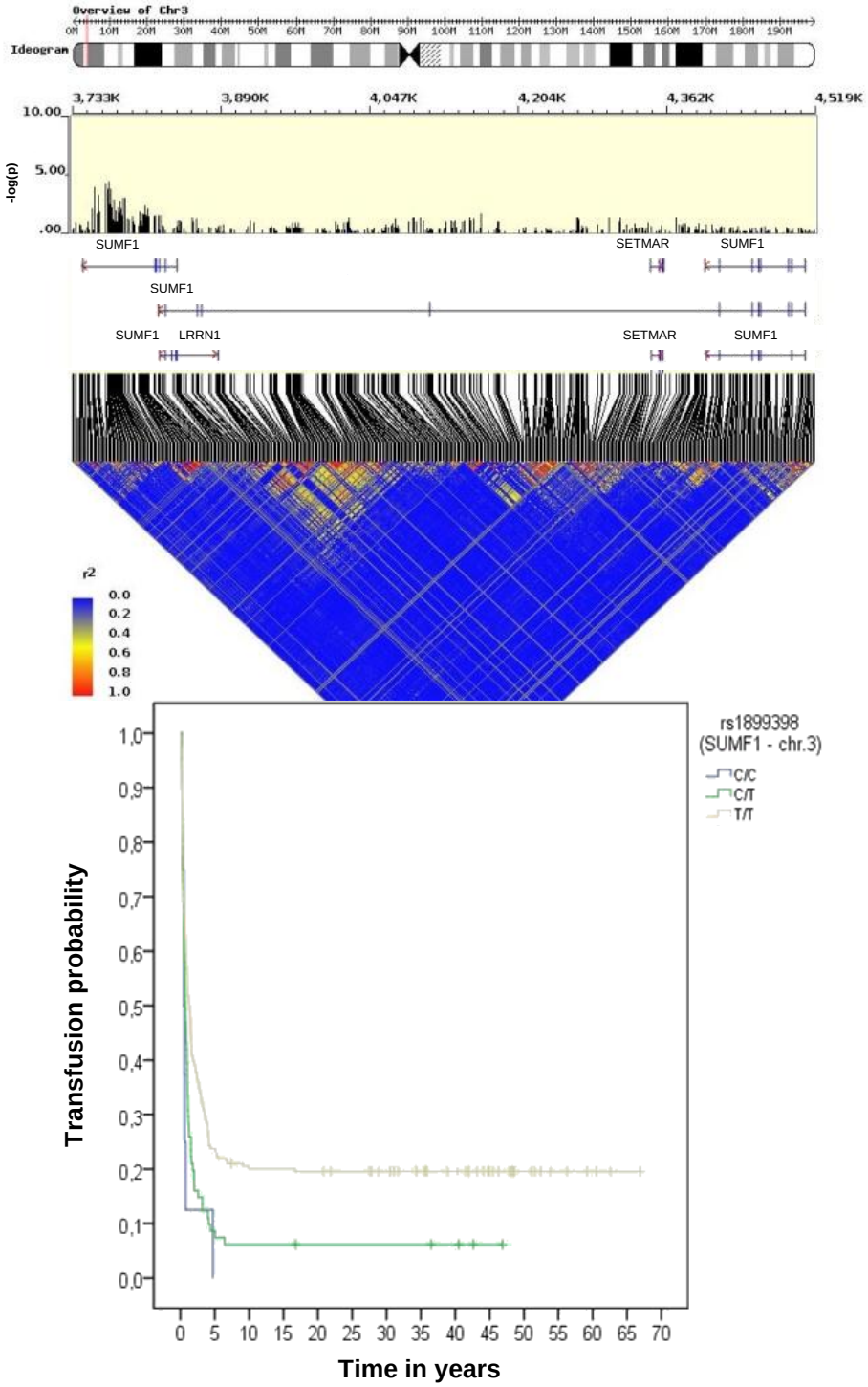


Figure 13. HBS1L-MYB intergenic region and effect of its respective most significant SNP on time to first transfusion.

[back to text]

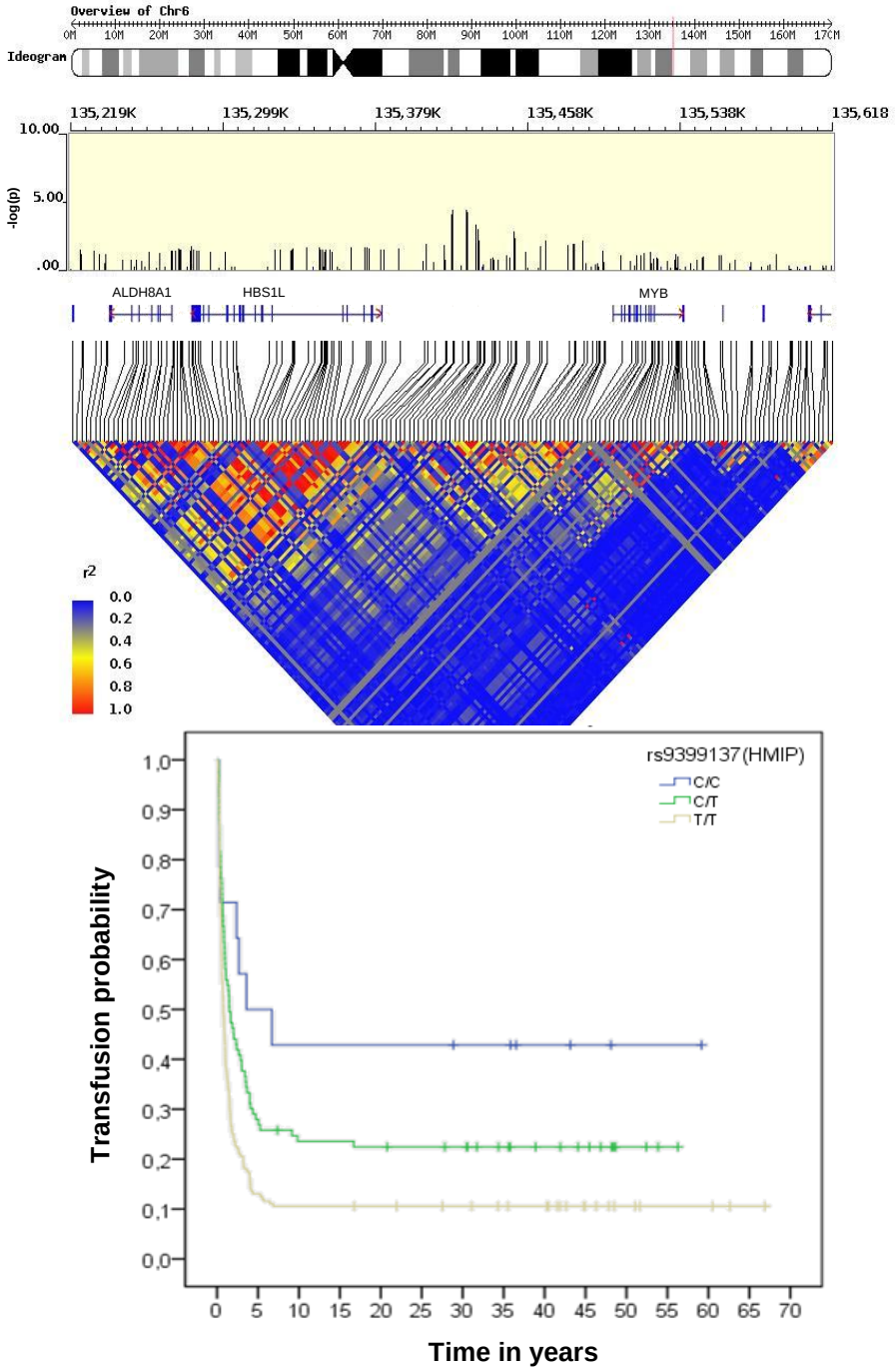


Figure 14. α-globin gene cluster and effect of α gene deletions on time to first transfusion.

[back to text]

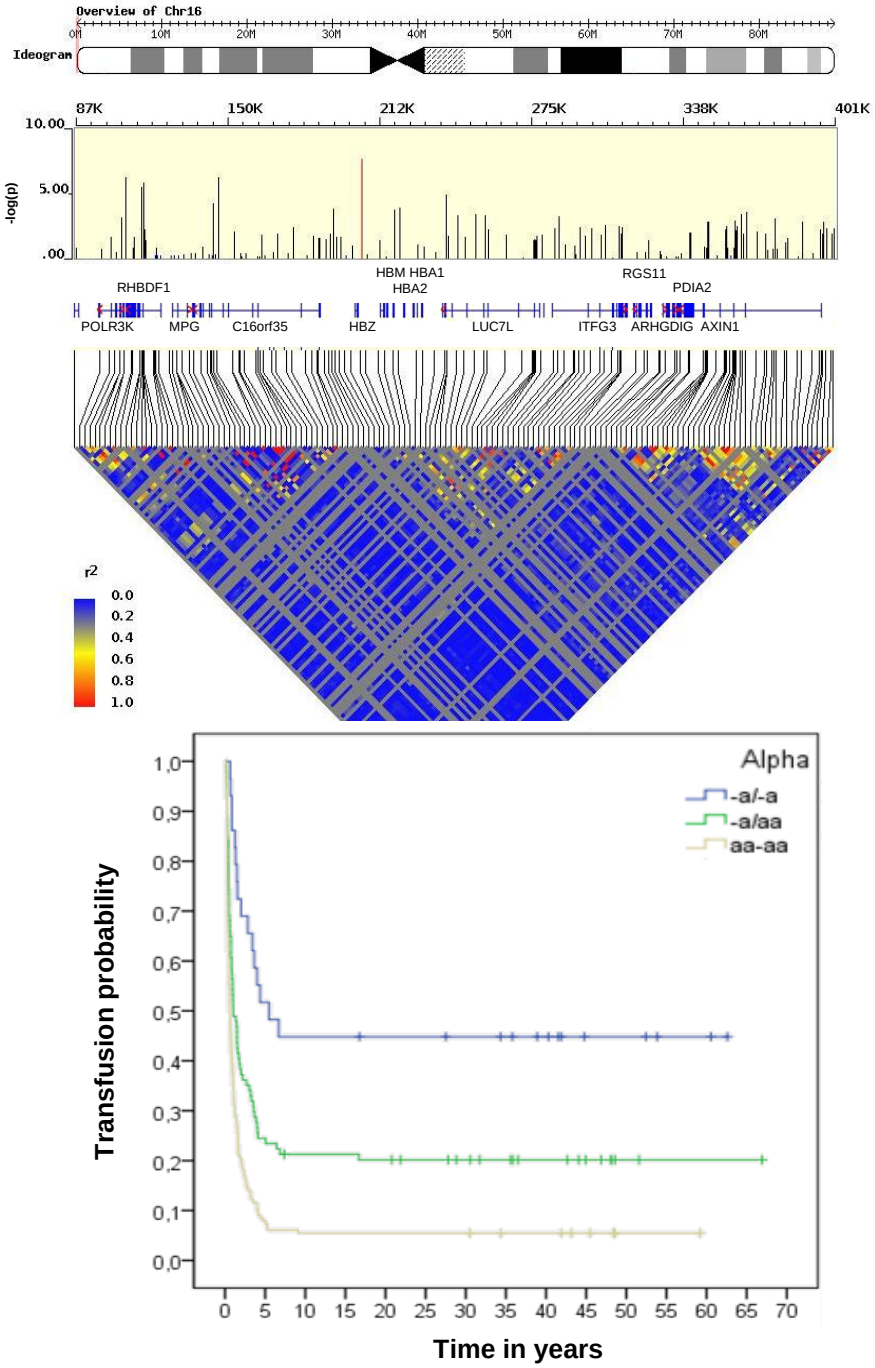


Figure 15. Survival functions for rs6600233 while accounting for gender, BCL11A, HBS1L-MYB intergenic region and α gene deletions effects on time to first transfusion, showing a significant difference between curves (p=0.007, exp(β)=1.33).

[back to text]

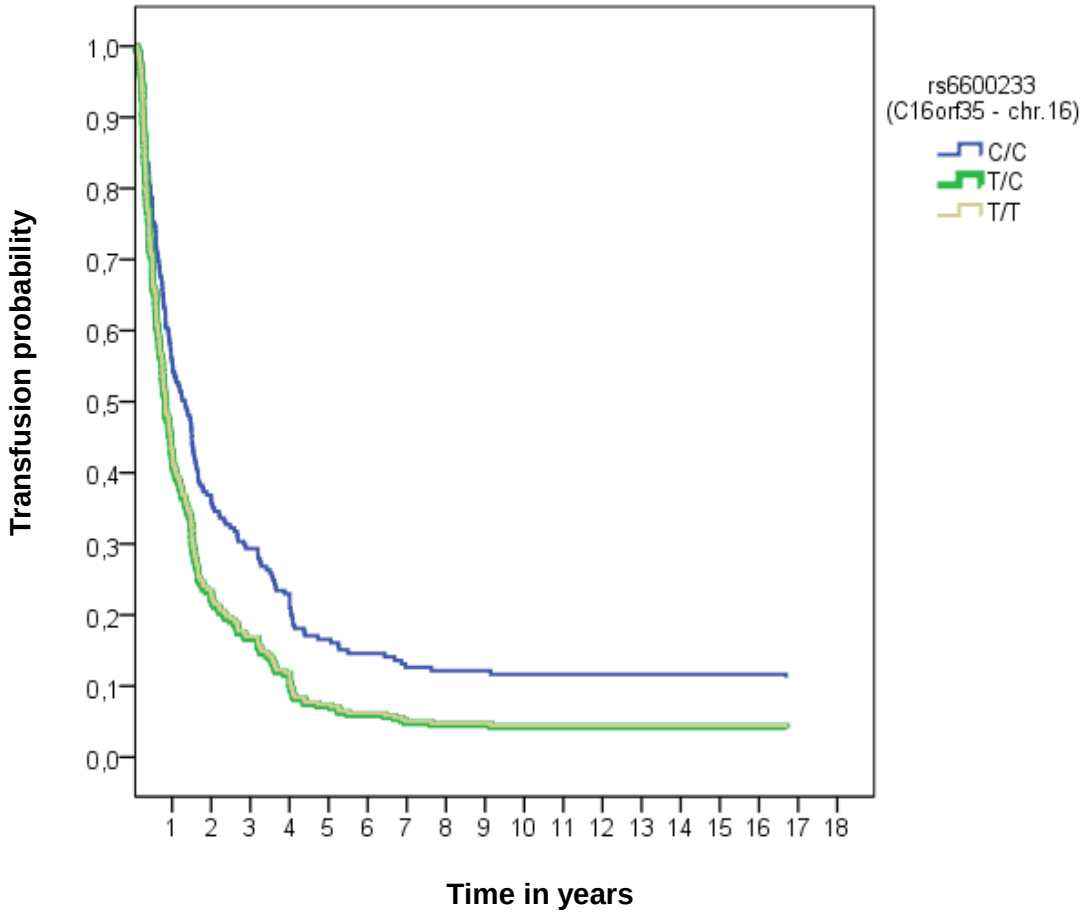
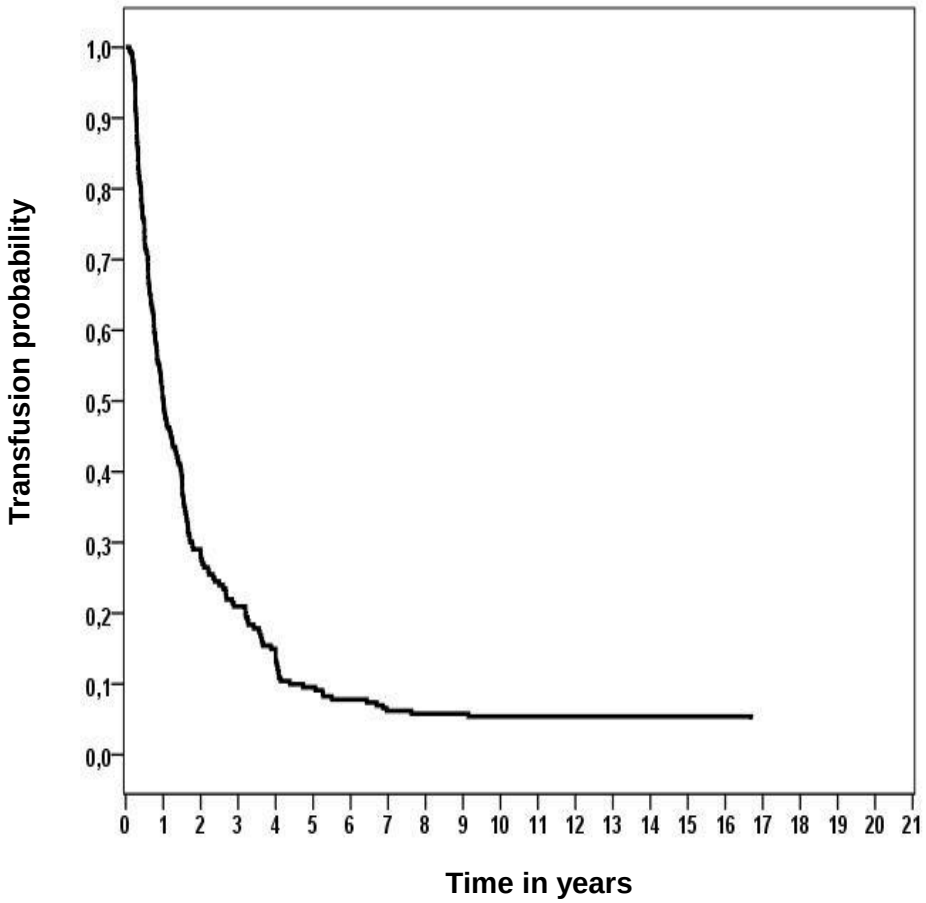


Figure 16. Survival function at the mean of covariates while accounting for effects on time to first transfusion of gender, BCL11A, HBS1L-MYB intergenic region, α gene deletions and suggestively significant results from the present study.

[back to text]



REFERENCES

1. Vichinsky, E.P. Changing patterns of thalassemia worldwide. *Ann. N. Y. Acad. Sci* 1054, 18-24 (2005).

2. Flint, J., Harding, R.M., Boyce, A.J. & Clegg, J.B. The population genetics of the haemoglobinopathies. *Baillieres Clin. Haematol* 11, 1-51 (1998).

3. Cao, A. & Galanello, R. Beta-thalassemia. *Genet. Med* 12, 61-76 (2010).

4. Weatherall, D.J. & Clegg, J.B. Inherited haemoglobin disorders: an increasing global health problem. *Bull World Health Organ* 79, 704-712 (2001).

5. Borgna-Pignatti, C. et al. Survival and complications in patients with thalassemia major treated with transfusion and deferoxamine. *Haematologica* 89, 1187-1193 (2004).

6. Huisman, T., Carver, M. & Efremov, G. A Syllabus of Human Hemoglobin Variants (1996). *The Sickle Cell Anemia Foundation, Augusta, GA, USA.*

7. Hardison, R.C. et al. Access to a syllabus of human hemoglobin variants (1996) via the World Wide Web. *Hemoglobin* 22, 113-127 (1998).

8. Hardison, R., Riemer, C., Chui, D.H., Huisman, T.H. & Miller, W. Electronic access to sequence alignments, experimental results, and human mutations as an aid to studying globin gene regulation. *Genomics* 47, 429-437 (1998).

9. Stamatoyannopoulos, G. Molecular and cellular basis of hemoglobin switching. *Disorders of hemoglobin: Genetics, pathophysiology, and clinical management* (2001).

10. Mosca, A., Paleari, R., Leone, D. & Ivaldi, G. The relevance of hemoglobin F measurement in the diagnosis of thalassemias and related hemoglobinopathies. *Clin. Biochem* 42, 1797-1801 (2009).

11. Thein, S.L. et al. Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood* 95, 342-346 (2000).

12. Sampietro, M., Thein, S.L., Contreras, M. & Pazmany, L. Variation of HbF and F-cell number with the G-gamma Xmn I (C-T) polymorphism in normal individuals. *Blood* 79, 832-833 (1992).

13. Thein, S.L. & Craig, J.E. Genetics of Hb F/F cell variance in adults and heterocellular hereditary persistence of fetal hemoglobin. *Hemoglobin* 22, 401-414 (1998).

14. Thein, S.L. & Weatherall, D.J. A non-deletion hereditary persistence of fetal hemoglobin (HPFH) determinant not linked to the beta-globin gene complex. *Prog. Clin. Biol. Res* 316B, 97-111 (1989).

15. Cappellini, M.D., Fiorelli, G. & Bernini, L.F. Interaction between homozygous beta (0) thalassaemia and the Swiss type of hereditary persistence of fetal haemoglobin. *Br. J. Haematol* 48, 561-572 (1981).

16. Galanello, R. et al. Molecular analysis of beta zero-thalassemia intermedia in Sardinia. *Blood* 74, 823-827 (1989).

17. Ho, P.J., Hall, G.W., Luo, L.Y., Weatherall, D.J. & Thein, S.L. Beta-thalassaemia intermedia: is it possible consistently to predict phenotype from genotype? *British Journal of Haematology* 100, 70-78 (1998).

18. Altay, C., Huisman, T.H. & Schroeder, W.A. Another form of the hereditary persistence of fetal hemoglobin (the Atlanta type)? *Hemoglobin* 1, 125-133 (1976).

19. Stamatoyannopoulos, G., Wood, W., Papayannopoulou, T. & Nute, P. A new form of hereditary persistence of fetal hemoglobin in blacks and its association with sickle cell trait. *Blood* 46, 683-692 (1975).

20. Gianni, A.M. et al. A gene controlling fetal hemoglobin expression in adults is not linked to the non-alpha globin cluster. *EMBO J* 2, 921-925 (1983).

21. Martinez, G., Novelletto, A., Di Rienzo, A., Felicetti, L. & Colombo, B. A case of hereditary persistence of fetal hemoglobin caused by a gene not linked to the beta-globin cluster. *Hum. Genet* 82, 335-337 (1989).

22. Giampaolo, A. et al. Heterocellular hereditary persistence of fetal hemoglobin (HPFH). Molecular mechanisms of abnormal gamma-gene expression in association with beta thalassemia and linkage relationship with the beta-globin gene cluster. *Hum. Genet* 66, 151-156 (1984).

23. Old, J.M., Ayyub, H., Wood, W.G., Clegg, J.B. & Weatherall, D.J. Linkage analysis of nondeletion hereditary persistence of fetal hemoglobin. *Science* 215, 981-982 (1982).

24. Menzel, S. et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet* 39, 1197-1199 (2007).

25. Galarneau, G. et al. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat Genet* 42, 1049-1051 (2010).

26. Orkin, S.H. et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science (New York, N.Y.)* 322, 1839-1842 (2008).

27. Uda, M. et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proceedings of the National Academy of Sciences of the United States of America* 105, 1620-1625 (2008).

28. Sedgewick, A.E. et al. BCL11A is a major HbF quantitative trait locus in three different populations with beta-hemoglobinopathies. *Blood Cells, Molecules & Diseases* 41, 255-258 (2008).

29. Close, J. et al. Genome annotation of a 1.5 Mb region of human chromosome 6q23 encompassing a quantitative trait locus for fetal hemoglobin expression in adults. *BMC Genomics* 5, 33 (2004).

30. Menzel, S. et al. The HBS1L-MYB intergenic region on chromosome 6q23.3 influences erythrocyte, platelet, and monocyte counts in humans. *Blood* 110, 3624-3626 (2007).

31. Thein, S.L. et al. Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proceedings of the National Academy of Sciences of the United States of America* 104, 11346-11351 (2007).

32. Lettre, G. et al. DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl. Acad. Sci. U.S.A* 105, 11869-11874 (2008).

33. Pandit, R. et al. Association of SNP in exon 1 of HBS1L with hemoglobin F level in β0-thalassemia/hemoglobin E. *International Journal of Hematology* 88, 357-361 (2008).

34. Wahlberg, K. et al. The HBS1L-MYB intergenic interval associated with elevated HbF levels shows characteristics of a distal regulatory region in erythroid cells. *Blood* 114, 1254-1262 (2009).

35. Haj Khelil, A. et al. Xmn I polymorphism associated with concomitant activation of (G)γ and (A)γ globin gene transcription on a β(0)-thalassemia chromosome. *Blood Cells Mol Dis* (2010).doi:10.1016/j.bcmd.2010.11.002

36. Sampietro, M., Thein, S.L., Contreras, M. & Pazmany, L. Variation of HbF and F-cell number with the G-gamma Xmn I (C-T) polymorphism in normal individuals. *Blood* 79, 832-833 (1992).

37. Gilman, J. & Huisman, T. DNA sequence variation associated with elevated fetal G gamma globin production. *Blood* 66, 783-787 (1985).

38. Thein, S.L. et al. Detection of a major gene for heterocellular hereditary persistence of fetal hemoglobin after accounting for genetic modifiers. *Am J Hum Genet* 54, 214-228 (1994).

39. Labie, D. et al. The -158 site 5' to the G gamma gene and G gamma expression. *Blood* 66, 1463-1465 (1985).

40. Labie, D. et al. Common haplotype dependency of high G gamma-globin gene expression and high Hb F levels in beta-thalassemia and sickle cell anemia patients. *Proc Natl Acad Sci U S A* 82, 2111-2114 (1985).

41. Garner, C. et al. Haplotype mapping of a major quantitative-trait locus for fetal hemoglobin production, on chromosome 6q23. *Am. J. Hum. Genet* 62, 1468-1474 (1998).

42. Creary, L.E. et al. Genetic variation on chromosome 6 influences F cell levels in healthy individuals of African descent and HbF levels in sickle cell patients. *PLoS ONE* 4, e4218 (2009).

43. Chen, Z., Luo, H., Steinberg, M.H. & Chui, D.H. BCL11A represses HBG transcription in K562 cells. *Blood Cells, Molecules, and Diseases* 42, 144-149 (2009).

44. Jawaid, K., Wahlberg, K., Thein, S.L. & Best, S. Binding patterns of BCL11A in the globin and GATA1 loci and characterization of the BCL11A fetal hemoglobin locus. *Blood Cells Mol Dis* (2010).doi:10.1016/j.bcmd.2010.05.006

45. Sankaran, V.G., Xu, J. & Orkin, S.H. Transcriptional silencing of fetal hemoglobin by BCL11A. *Ann. N. Y. Acad. Sci* 1202, 64-68 (2010).

46. Xu, J. et al. Transcriptional silencing of γ-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev* 24, 783-798 (2010).

47. Sankaran, V.G. et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science* 322, 1839-1842 (2008).

48. Jiang, J. et al. cMYB is involved in the regulation of fetal hemoglobin production in adults. *Blood* 108, 1077-1083 (2006).

49. Dover, G.J. et al. Fetal hemoglobin levels in sickle cell disease and normal individuals are partially controlled by an X-linked gene located at Xp22. 2. *Blood* 80, 816 (1992).

50. Miyoshi, K. et al. X-linked dominant control of F-cells in normal adult life: characterization of the Swiss type as hereditary persistence of fetal hemoglobin regulated dominantly by gene(s) on X chromosome. *Blood* 72, 1854-1860 (1988).

51. Garner, C. et al. Interaction between two quantitative trait loci affects fetal haemoglobin expression. *Ann. Hum. Genet* 69, 707-714 (2005).

52. Garner, C. et al. Quantitative trait locus on chromosome 8q influences the switch from fetal to adult hemoglobin. *Blood* 104, 2184-2186 (2004).

53. Garner, C.P., Tatu, T., Best, S., Creary, L. & Thein, S.L. Evidence of genetic interaction between the beta-globin complex and chromosome 8q in the expression of fetal hemoglobin. *Am. J. Hum. Genet* 70, 793-799 (2002).

54. Origa, R. et al. Cholelithiasis in thalassemia major. *Eur. J. Haematol* 82, 22-25 (2009).

55. Weatherall, D.J. Pathophysiology of thalassaemia. *Baillieres Clin. Haematol* 11, 127-146 (1998).

56. Singer, S.T. & Ataga, K.I. Hypercoagulability in sickle cell disease and beta-thalassemia. *Curr. Mol. Med* 8, 639-645 (2008).

57. Taher, A.T., Otrock, Z.K., Uthman, I. & Cappellini, M.D. Thalassemia and hypercoagulability. *Blood Reviews* 22, 283-292 (2008).

58. Kulozik, A. et al. Fetal hemoglobin levels and beta s globin haplotypes in an Indian populations with sickle cell disease. *Blood* 69, 1742-1746 (1987).

59. Uda, M. et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc. Natl. Acad. Sci. U.S.A* 105, 1620-1625 (2008).

60. Solovieff, N. et al. Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. *Blood* (2009).doi:10.1182/blood-2009-08-239517

61. Nuinoon, M. et al. A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. *Hum. Genet* 127, 303-314 (2010).

62. Thein, S.L. et al. Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl. Acad. Sci. U.S.A* 104, 11346-11351 (2007).

63. Gauderman WJ, Morrison JM QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. (2006).

64. Wang, K. et al. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* 17, 1665 - 1674 (2007).

65. Diskin, S.J. et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research* 36, e126 (2008).

66. Wang, K. et al. Modeling genetic inheritance of copy number variations. *Nucleic Acids Research* 36, e138 (2008).

67. Colella, S. et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 35, 2013-2025 (2007).

68. Purcell, S. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 81, 559-575 (2007).

69. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol* 34, 816-834 (2010).

70. The International HapMap Project. *Nature* 426, 789-796 (2003).

71. Team, R.D.C. R: A Language and Environment for Statistical Computing. 1, ISBN 3-900051-07-0

72. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904-909 (2006).

73. Daly, M. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 81, 559-575 (2007).

74. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. GRR: graphical representation of relationship errors. *Bioinformatics* 17, 742-743 (2001).

75. Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet* 76, 887-893 (2005).

76. Winchester, L., Yau, C. & Ragoussis, J. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic* 8, 353-366 (2009).

77. Hoggart, C.J. et al. Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet* 72, 1492-1504 (2003).

78. Barnholtz-Sloan, J.S., Chakraborty, R., Sellers, T.A. & Schwartz, A.G. Examining population stratification via individual ancestry estimates versus self-reported race. *Cancer Epidemiol. Biomarkers Prev* 14, 1545-1551 (2005).

79. Reiner, A.P. et al. Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study. *Am. J. Hum. Genet* 76, 463-477 (2005).

80. Yang, N. et al. Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Hum. Genet* 118, 382-392 (2005).

81. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904-909 (2006).

82. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet* 67, 170-181 (2000).

83. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* 55, 997-1004 (1999).

84. GenGen website. <http://www.openbioinformatics.org/gengen/index.html>

85. Diskin, S.J. et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res* 36, e126 (2008).

86. Marioni, J.C. et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol* 8, R228 (2007).

87. Ge, D. et al. WGAViewer: Software for genomic annotation of whole genome association studies. *Genome Res* 18, 640-643 (2008).

88. S original by Terry Therneau and ported by Thomas Lumley survival: Survival analysis, including penalised likelihood. *R package version 2.9.2*

89. Hubbard, T.J.P. et al. Ensembl 2007. *Nucleic Acids Research* 35, D610-D617 (2007).

90. PubMed. <http://www.ncbi.nlm.nih.gov/pubmed>

91. Yu, W., Clyne, M., Khoury, M.J. & Gwinn, M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* 26, 145 - 146 (2010).

92. Hindorff, L.A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A* 106, 9362-9367 (2009).

93. Ganesh, S.K. et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet* 41, 1191-1198 (2009).

94. Kamatani, Y. et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet* 42, 210-215 (2010).

95. Ma, Q. et al. Fetal hemoglobin in sickle cell anemia: genetic determinants of response to hydroxyurea. *Pharmacogenomics J* 7, 386-394 (2007).

96. Sherva, R. et al. Genetic modifiers of Hb E/beta0 thalassemia identified by a two-stage genome-wide association study. *BMC Medical Genetics* 11, 51 (2010).

97. Galanello, R. et al. Amelioration of Sardinian {beta}0 thalassemia by genetic modifiers. *Blood* 114, 3935-3937 (2009).

98. Romao, L., Osorio-Almeida, L., Higgs, D.R., Lavinha, J. & Liebhaber, S.A. Alpha-thalassemia resulting from deletion of regulatory sequences far upstream of the alpha-globin structural genes. *Blood* 78, 1589-1595 (1991).

99. De Gobbi, M. et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312, 1215-1217 (2006).

100. Ribeiro, D.M., Zaccariotto, T.R., Santos, M.N.N., Costa, F.F. & Sonati, M.F. Influence of the polymorphisms of the alpha-major regulatory element HS-40 on in vitro gene expression. *Braz. J. Med. Biol. Res* 42, 783-786 (2009).

101. Selig, S., Lidov, H.G., Bruno, S.A., Segal, M.M. & Kunkel, L.M. Molecular characterization of Br-cadherin, a developmentally regulated, brain-specific cadherin. *Proc. Natl. Acad. Sci. U.S.A* 94, 2398-2403 (1997).

102. Bankovic, J. et al. Identification of genes associated with non-small-cell lung cancer promotion and progression. *Lung Cancer* 67, 151-159 (2010).

103. Chalmers, I.J., Höfler, H. & Atkinson, M.J. Mapping of a Cadherin Gene Cluster to a Region of Chromosome 5 Subject to Frequent Allelic Loss in Carcinoma. *Genomics* 57, 160-163 (1999).

104. Settembre, C. et al. A block of autophagy in lysosomal storage disorders. *Hum. Mol. Genet* 17, 119-129 (2008).

105. Bernales, S., McDonald, K.L. & Walter, P. Autophagy counterbalances endoplasmic reticulum expansion during the unfolded protein response. *PLoS Biol* 4, e423 (2006).

106. Yorimitsu, T. & Klionsky, D.J. Eating the endoplasmic reticulum: quality control by autophagy. *Trends Cell Biol* 17, 279-285 (2007).

107. Yorimitsu, T., Nair, U., Yang, Z. & Klionsky, D.J. Endoplasmic reticulum stress triggers autophagy. *J. Biol. Chem* 281, 30299-30304 (2006).

108. Ding, W. et al. Linking of autophagy to ubiquitin-proteasome system is important for the regulation of endoplasmic reticulum stress and cell viability. *Am. J. Pathol* 171, 513-524 (2007).

109. Schroder, M. & Kaufman, R.J. The mammalian unfolded protein response. *Annu. Rev. Biochem.* 74, 739-789 (2005).

110. Buono, M., Visigalli, I., Bergamasco, R., Biffi, A. & Cosma, M.P. Sulfatase modifying factor 1-mediated fibroblast growth factor signaling primes hematopoietic multilineage development. *J. Exp. Med* 207, 1647-1660 (2010).

111. Basu, S., Banerjee, D., Chandra, S. & Chakrabarti, A. Eryptosis in hereditary spherocytosis and thalassemia: role of glycoconjugates. *Glycoconj. J* 27, 717-722 (2010).

112. Greco, F. et al. Rap2, but not Rap1 GTPase is expressed in human red blood cells and is involved in vesiculation. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1763, 330-335 (2006).

113. Torti, M. et al. Interaction of the low-molecular-weight GTP-binding protein rap2 with the platelet cytoskeleton is mediated by direct binding to the actin filaments. *J. Cell. Biochem* 75, 675-685 (1999).

114. Eldor, A. & Rachmilewitz, E.A. The hypercoagulable state in thalassemia. *Blood* 99, 36-43 (2002).

115. Alekperova, G.A., Orudzhev, A.G. & Javadov, S.A. Analysis of erythrocyte and platelet membrane proteins in various forms of beta-thalassemia. *Biochemistry Mosc* 69, 748-753 (2004).