DNA polymorphisms at the *BCL11A*, *HBS1L-MYB*, and *β-globin* loci associate with fetal hemoglobin levels and pain crises in sickle cell disease

Guillaume Lettre*[†‡§], Vijay G. Sankaran[‡¶‖], Marcos André C. Bezerra**, Aderson S. Araújo**, Manuela Uda[††], Serena Sanna[††], Antonio Cao[††], David Schlessinger[‡‡], Fernando F. Costa[§§], Joel N. Hirschhorn*[†§], and Stuart H. Orkin[§¶‖]

*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142; [†]Divisions of Genetics and Endocrinology and Program in Genomics and [¶]Division of Hematology/Oncology, Children's Hospital Boston, Boston, MA 02115; [‖]Department of Pediatric Oncology, Dana–Farber Cancer Institute, Boston, MA 02115; **Fundação de Hematologia e Hemoterapia de Pernambuco, Hemope, 52011-000 Recife, Brazil; [††]Istituto di Neurogenetica e Neurofarmacologia (INN), Consiglio Nazionale delle Ricerche, c/o Cittadella Universitaria di Monserrato, Monserrato, 09042 Cagliari, Italy; [‡‡]Gerontology Research Center, National Institute on Aging, 5600 Nathan Shock Drive, Baltimore, MD 21224; and [§§]Center of Hemotherapy and Hematology, State University of Campinas, 13083-970 Campinas, Sau Paulo, Brazil

Contributed by Stuart H. Orkin, May 19, 2008 (sent for review April 14, 2008)

Sickle cell disease (SCD) is a debilitating monogenic blood disorder with a highly variable phenotype characterized by severe pain crises, acute clinical events, and early mortality. Interindividual variation in fetal hemoglobin (HbF) expression is a known and potentially heritable modifier of SCD severity. High HbF levels are correlated with reduced morbidity and mortality. Common single nucleotide polymorphisms (SNPs) at the *BCL11A* and *HBS1L-MYB* loci have been implicated previously in HbF level variation in nonanemic European populations. We recently demonstrated an association between a *BCL11A* SNP and HbF levels in one SCD cohort [Uda M, *et al.* (2008) *Proc Natl Acad Sci* USA 105:1620–1625]. Here, we genotyped additional *BCL11A* SNPs, *HBS1L-MYB* SNPs, and an SNP upstream of $^{G}\gamma$*-globin* (*HBG2*; the *Xmn*I polymorphism), in two independent SCD cohorts: the African American Cooperative Study of Sickle Cell Disease (CSSCD) and an SCD cohort from Brazil. We studied the effect of these SNPs on HbF levels and on a measure of SCD-related morbidity (pain crisis rate). We strongly replicated the association between these SNPs and HbF level variation (in the CSSCD, *P* values range from 0.04 to 2×10^{-42}). Together, common SNPs at the *BCL11A*, *HBS1L-MYB*, and *β-globin* (*HBB*) loci account for >20% of the variation in HbF levels in SCD patients. We also have shown that HbF-associated SNPs associate with pain crisis rate in SCD patients. These results provide a clear example of inherited common sequence variants modifying the severity of a monogenic disease.

genetic modifier | single nucleotide polymorphism | globin gene regulation

Sickle cell disease (SCD) is a Mendelian disorder caused by a point mutation leading to a single amino acid substitution in the beta subunit of hemoglobin, the principal oxygen transporter in red blood cells (RBCs). It has been estimated that SCD results in the annual loss of several millions of disability-adjusted life years, particularly in the developing world (1). Clinical severity of SCD is extremely variable, and the reasons for this heterogeneity are not fully understood. A notable feature of SCD is the frequent occurrence of pain and acute clinical episodes, which are generally attributable to vaso-occlusive crises (2) and the acute chest syndrome (ACS) of SCD (3). As a consequence of these complications, SCD patients have increased mortality as compared with control populations (4). Although the risk factors underlying these complications are not well characterized, higher expression of fetal hemoglobin (HbF) in adulthood ameliorates morbidity and mortality in SCD (2–5). Thus, interindividual variation in HbF levels is likely one of the main modifiers that contribute to the clinical heterogeneity observed in SCD patients.

Extensive observations on the natural history of SCD have led to efforts to stimulate HbF expression to treat patients. Experimental and clinical work in this area has led to the use of hydroxyurea, an agent that was found empirically to increase the production of HbF (6, 7). Because HbF levels vary considerably between individuals (particularly in populations with certain anemias) and are a highly heritable trait ($h^2 > 0.7$) (8), it is expected that the identification of genetic polymorphisms that modulate HbF levels will shed light on the molecular mechanisms that control HbF expression and on the etiology of the clinical heterogeneity observed in SCD patients. Such insights might ultimately identify novel drug targets for new treatments of SCD.

Recently, genetic association studies have pinpointed several single nucleotide polymorphisms (SNPs) that are reproducibly associated with variation in the expression of HbF into adulthood in healthy European populations (9–11). These SNPs are located in the gene *BCL11A* on chromosome 2 and in the *HBS1L-MYB* intergenic region on chromosome 6. These initial discoveries were made in nonanemic populations (Northern Europeans and Sardinians), in whom the level of HbF expression is generally <1%. Uda and colleagues also examined the clinical implications of their findings in an independent β-thalassemia cohort: the high-HbF allele at the *BCL11A* SN*P* was significantly more frequent in patients with milder clinical forms of β-thalassemia, suggesting that this genetic polymorphism may be an important genetic modulator of disease severity (11). In the same study, we reported replication of the association between a *BCL11A* SNP and HbF levels in the African American Cooperative Study of Sickle Cell Disease (CSSCD) (11).

Here, we replicate the effect that SNPs at the *BCL11A* locus have on HbF levels in SCD patients in an independent cohort from Brazil. In addition, we demonstrate a clear association between SNPs in the *HBS1L-MYB* intergenic region and HbF levels in both the CSSCD and the Brazilian SCD cohort. Finally, we show that a panel of five SNPs from three loci (*BCL11A*,

Author contributions: G.L., V.G.S., J.N.H., and S.H.O. designed research; G.L. and V.G.S. performed research; M.A.C.B., A.S.A., M.U., S.S., A.C., D.S., and F.F.C. contributed new reagents/analytic tools; G.L., V.G.S., J.N.H., and S.H.O. analyzed data; and G.L., V.G.S., J.N.H., and S.H.O. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

See Commentary on page 11595.

[‡]G.L. and V.G.S. contributed equally to this work.

[§]To whom correspondence may be addressed. E-mail: lettre@broad.mit.edu, joelh@broad.mit.edu, or stuart_orkin@dfci.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0804799105/DCSupplemental.

Table 1. Fetal hemoglobin association results for SNPs at the *BCL11A*, *HBS1L-MYB*, and *β-globin* loci in the CSSCD and the Brazil sickle cell disease cohort

Gene and chromosome (physical position)*	SNP	CSSCD (N = 1,275) MAF (allele)†	Effect size (SE)‡	Variance explained (%)	P value	Brazil (N = 350) MAF (allele)†	Effect size (SE)‡	Variance explained (%)	P value
BCL11A									
2 (60573750)	rs11886868	0.31 (C)	0.524 (0.041)	11.8	4×10^{-35}	0.39 (C)	0.376 (0.078)	6.7	2×10^{-6}
2 (60574475)	rs4671393	0.27 (A)	0.598 (0.042)	14.1	2×10^{-42}	0.24 (A)	0.485 (0.086)	9.0	3×10^{-8}
2 (60574851)	rs7557939	0.31 (G)	0.540 (0.040)	12.6	6×10^{-38}	0.38 (G)	0.409 (0.076)	7.9	1×10^{-7}
HBS1L-MYB									
6 (135417902)	rs28384513	0.20 (C)	−0.102 (0.049)	0.4	0.04	Not genotyped			
6 (135460609)	rs7776054	0.20 (G)	0.103 (0.050)	0.4	0.04	0.17 (G)	0.346 (0.103)	3.4	0.0009
6 (135460711)	rs9399137	0.06 (C)	0.571 (0.086)	3.5	5×10^{-11}	Failed			
6 (135461324)	rs9389268	0.19 (G)	0.102 (0.051)	0.3	0.05	0.20 (G)	0.349 (0.093)	4.0	0.0002
6 (135468266)	rs4895441	0.10 (G)	0.338 (0.064)	2.2	1×10^{-7}	0.14 (G)	0.552 (0.107)	7.4	4×10^{-7}
β-globin locus									
11 (5232745)	rs7482144	0.07 (A)	0.407 (0.080)	2.2	4×10^{-7}	Homozygous			

*Position on NCBI Build 36.1.
†MAF, minor allele frequency. Minor alleles (positive strand) are given in the parentheses.
‡Effect sizes and standard errors are given in standard deviation units for the minor allele.

HBS1L-MYB, and the *β-globin* gene cluster), which together account for >20% of variation in HbF levels, associate with SCD-related pain crises in the large prospective CSSCD. Our findings hold potential for increased biological understanding of the clinical variability in SCD, with possible implications for future drug development. In addition, after further validation in additional prospective patient cohorts, these and other such results could eventually help guide the development of treatment plans tailored for different levels of predicted risk of SCD-related complications.

Results

Association of Genetic Variants to HbF Levels. The association between the *Xmn*I polymorphism (rs7482144) in the proximal promoter of the $^{G}\gamma$-*globin* (*HBG2*) gene and HbF levels is well documented in SCD patients (12, 13). In an effort to collect genotypes for all variants robustly associated with HbF levels, we genotyped rs7482144 in the CSSCD and clearly confirmed the previously described association ($P = 4 \times 10^{-7}$; Table 1 and Fig. 1). In the CSSCD, rs7482144 explains 2.2% of the variation in HbF levels. We could not test this association in the Brazil SCD cohort because rs7482144 was monomorphic (patients in this Brazilian population have neither the Senegal nor the Arab-Indian sickle cell haplotype that contain the *Xmn*I variant).

Genome-wide linkage scans identified a significant F-cell linkage peak at 6q23 in a large Asian–Indian kindred with β-thalassemia and persistence of HbF expression (14, 15). This linkage signal was refined to the intergenic region between genes *HBS1L* and *MYB* by using genetic association analysis in healthy northern European twins (9). Using stepwise regression, Thein and colleagues concluded that at least three independent genetic variants (rs28384513, rs9399137, and rs6929404) explain the linkage peak observed at 6q23 (9). More recently, an association signal to HbF levels in the *HBS1L-MYB* intergenic region was also identified in a large nonanemic Sardinian cohort (10). We genotyped SNPs rs28384513 and rs9399137 identified by Thein *et al.* (using the iPLEX genotyping platform, we could not efficiently genotype rs6929404), as well as three SNPs reported by the Sardinia study (rs7776054, rs9389268, and rs4895441) in the CSSCD and Brazil SCD cohorts (Table 1). The association signal to HbF levels could be convincingly replicated in both panels, thus establishing that genetic variants at the *HBS1L-MYB* locus associate with HbF in SCD patients.

We considered whether the multiple variants at the *HBS1L-MYB* locus represented independent signals of association. Using stepwise regression, we found that rs28384513, rs9399137, and rs4895441 are independent association signals in the CSSCD data set (P values were 0.0002, 8×10^{-7}, and 0.0004 for rs28384513, rs9399137, and rs4895441 in the multivariate regression model, respectively), together explaining 5.0% of the variation in HbF levels [Table 1, Fig. 1, and supporting information (SI) Table S1]. Interestingly, the correlation due to linkage disequilibrium (LD) between each of these three variants in the CSSCD cohort was low ($r^2 < 0.2$), whereas rs9399137 and rs4895441 show strong LD in the HapMaP Northern European (CEU) samples ($r^2 = 0.8$) but not in the HapMaP African (YRI) samples ($r^2 = 0.004$); these results are consistent with lower degrees of LD often seen in individuals with recent African ancestry. Although the finding that rs28384513 and rs9399137 are independent HbF-associated SNPs had been reported previously (9), rs4895441 is a new "independent" genetic variant associated to variation in HbF levels at the *HBS1L-MYB* locus. The strong correlations present in populations of recent European ancestry had prevented resolution of this additional independent signal in prior studies (9). For the analyses of disease complications presented below, we tested the three independent *HBS1L-MYB* HbF-associated SNPs (rs28384513, rs9399137, and rs4895441) as predictive variables because they provide nonredundant information in the CSSCD.

More recently, common SNPs in the *BCL11A* gene on chromosome 2 have been shown to associate with F-cell or HbF levels in nonanemic northern European twins and Sardinians in genome-wide association studies (10, 11). To extend this finding, we have shown that an intronic SNP in *BCL11A*, rs11886868, strongly correlates with HbF levels in the CSSCD (ref 11 and Table 1). We genotyped rs11886868 in the SCD cohort from Brazil and could strongly replicate the association (P value = 2×10^{-6}, Table 1). The Sardinian study also highlighted other SNPs in LD with rs11886868, which had highly significant associations with HbF levels. To refine the association signal, we genotyped two of these SNPs, rs4671393 and rs7557939, in the CSSCD and Brazil SCD cohorts (Table 1 and Fig. 1). Both of these SNPs were more strongly associated with HbF level variation than rs11886868. Strikingly, the fraction of HbF phenotypic variation explained by rs4671393 in the CSSCD and the Brazil cohort is, respectively, 14.1% and 9.0%. In both SCD cohorts, when we conditioned the association analysis on rs4671393, association at rs11886868 and rs7557939 became nonsignificant ($P > 0.05$),

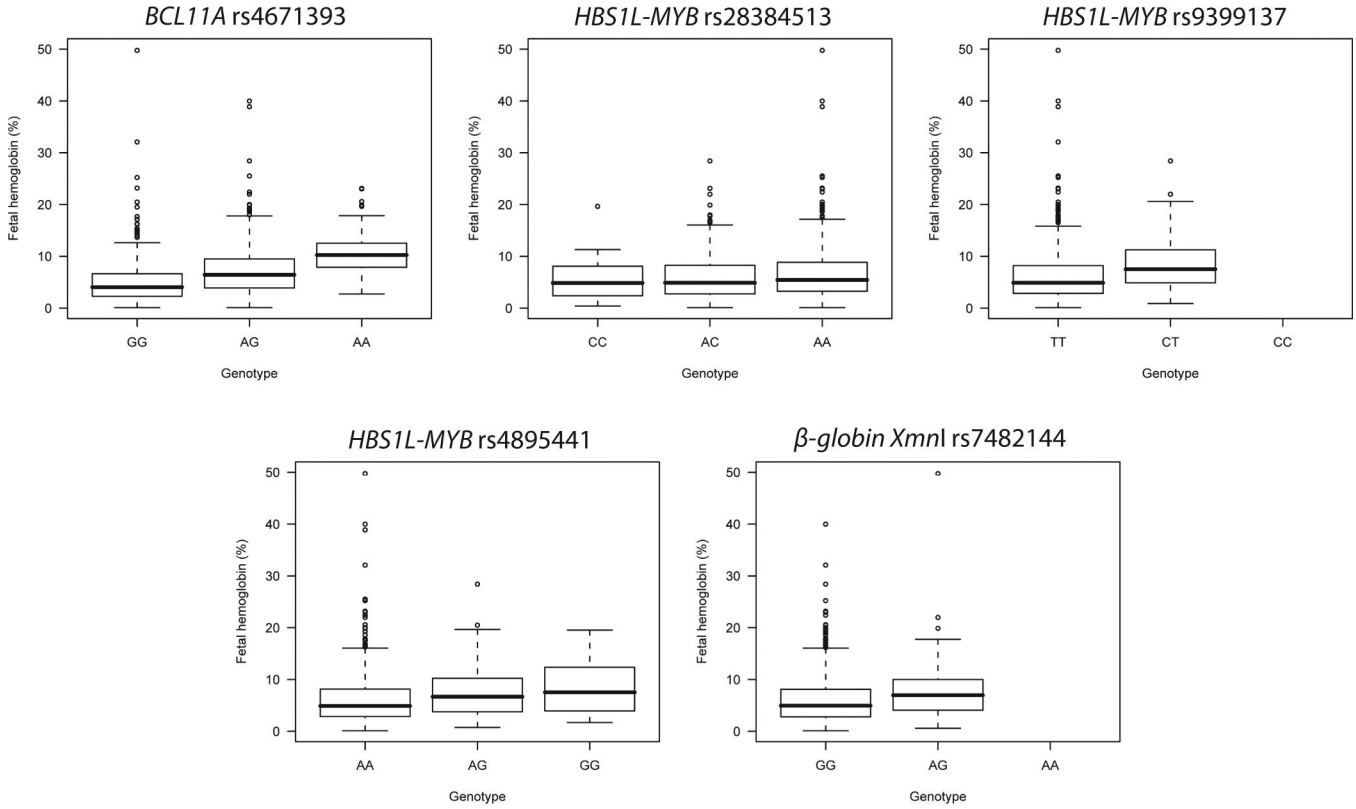

Fig. 1. Distribution of fetal hemoglobin levels conditioned on SNP genotypes. For *HBS1L-MYB* rs9399137 and *β-globin Xmn*I rs7482144, there are no individuals with the homozygote minor allele genotype. Boxes have lines at the lower quartile, median, and upper quartile. The plot whiskers extend up and down from the median a distance 1.5 times the interquartile range of the boxes (truncated at zero where necessary). Outliers are the points outside the whiskers indicated as circles.

suggesting that these markers tag the same causal polymorphism. All other possible conditional analyses with the three SNPs genotyped at the *BCL11A* locus confirmed that rs4671393 is, to date, the best known signal of association to HbF levels in SCD populations.

Effect of HbF-Associated Variants on Pain Crises in SCD. Previous analyses of the CSSCD dataset showed that increased HbF levels correlate with less severe complications [pain crises (2) and ACS (3)] and improved survival (4). Having shown that five sequence variants at three loci (*BCL11A* rs4671393, *HBS1L-MYB* rs28384513, rs9399137 and rs4895441, and *β-globin* gene cluster *Xmn*I rs7482144) robustly associate with HbF levels and explain >20% of the phenotypic variation in the trait (Fig. 1), we first asked whether these SNPs also correlate with pain crisis rate in SCD, and then whether they provide additional information beyond simple baseline HbF measurements. For this analysis, we used Poisson regression and fitted the same covariates (gender, age at entry, and hematocrit level) used in the models originally described by the CSSCD investigators (2).

Platt *et al.* (2) originally reported that steady-state HbF level is a strong predictor of the frequency of pain crises, a result that we could repeat (Table 2; pain HbF model vs. pain basic model, $P = 2 \times 10^{-5}$). We then replaced HbF level in the model by genotypes at one of the five HbF-associated SNPs, or the five SNPs altogether (Table 2; pain genotype model). For each SNP, adding an extra copy of the high-HbF allele reduced the pain rate, but the effect was not significant for any individual SNP (Table S2). However, the pain genotype model (which includes the five HbF-associated SNPs) was significant when compared to the pain basic model (Table 2; $P = 0.001$), indicating that markers associated with HbF levels, in aggregate, have power to predict pain crises in SCD. Finally, we tested the pain full model, in which HbF levels and HbF-associated SNP genotypes are included among the predictors, for association to pain rate: this model was significant when compared to the pain basic model

Table 2. HbF levels-associated SNPs and pain crisis in SCD

Model	χ^2 (degrees of freedom)	*P* value
Basic model: Pain rate = gender + age at entry + hematocrit level	n.a.	n.a.
HbF model: Pain rate = basic model + fetal hemoglobin level	Vs. basic model: 17.9 (1)	2×10^{-5}
Genotype model: Pain rate = basic model + rs28384513 + rs4671393 + rs9399137 + rs4895441 + rs7482144	Vs. basic model: 20.4 (5)	0.001
Full model: Pain rate = basic model + fetal hemoglobin level + rs28384513 + rs4671393 + rs9399137 + rs4895441 + rs7482144	Vs. basic model: 32.3 (6)	1×10^{-5}
	Vs. HbF model: 14.4 (5)	0.01

Chi square values are calculated by subtracting the deviance of the large model from the deviance of the smaller model. n.a., not applicable.

Table 3. Description of the sickle cell disease cohorts

Trait	CSSCD*	CSSCD (subset)†	Brazil
Number of males/ females	682 / 593	484 / 411	162 / 188
Age‡	14.5 ± 12.1	16.9 ± 11.4	n.a.
Follow-up (year)‡	6.5 ± 1.8	6.7 ± 1.0	11.4 ± 6.9
Fetal hemoglobin (%)‡	6.4 ± 4.7	6.1 ± 4.2	9.2 ± 6.0
Hemoglobin (g/dl)‡	8.6 ± 1.3	8.5 ± 1.2	7.5 ± 1.0
Hematocrit (%)‡	24.9 ± 4.1	24.6 ± 3.7	23.3 ± 3.0
Red blood cell count (x 10^{12}/L)‡	2.8 ± 0.6	2.8 ± 0.6	2.5 ± 0.6
MCV‡	89.4 ± 9.0	89.8 ± 8.6	94.4 ± 9.7
MCH‡	30.1 ± 2.9	30.2 ± 2.8	30.5 ± 3.7
White blood cell count (x 10^{9}/L)‡	11.9 ± 2.6	11.9 ± 2.7	13.0 ± 4.3
Monocyte (%)‡	7.4 ± 4.5	7.4 ± 4.3	8.7 ± 3.1
Platelet count (x 10^{9}/L)‡	442 ± 151	448 ± 154	412 ± 134
Pain rate (events/patient-year)‡	0.7 ± 1.4	0.9 ± 1.5	n.a.
Acute chest syndrome rate (events/patient-year)‡	0.1 ± 0.3	0.1 ± 0.2	n.a.
Number of deaths	45	37	13
Number of participants with renal complication	184	147	n.a.
Number of participants with seizure	31	30	n.a.

n.a., not available.

*Restricted to the CSSCD participants with DNA available for genotyping.

†Subset of CSSCD participants used in the analysis of SCD-related complications. See *Materials and Methods* for selection criteria.

‡Mean ± SD is given.

(Table 2, $P = 1 \times 10^{-5}$), and remained significant when compared to the pain HbF model (Table 2, $P = 0.01$). This latter result suggests that HbF-associated SNPs provide information on pain rate beyond their effect on single measurements of steady-state HbF levels.

We also analyzed the effect of these HbF-associated SNPs on ACS rate and survival using the same strategy as the CSSCD investigators (3, 4). We found that adding one or all of the HbF-associated SNPs as predictors of ACS rate or death did not improve the statistical models (Tables S2–S4). Because the numbers of ACS events ($n = 617$) and deaths ($n = 37$) are low in the set of CSSCD participants used in these analyses (Table 3), it would be premature to conclude that SNPs associated with HbF levels do not affect ACS and survival in SCD patients.

Discussion

The clinical heterogeneity of SCD presents a challenge in patient management. The clinical course of SCD can vary from an entirely benign one, in which patients may not even be aware of their disease status [particularly in certain SCD populations with very high levels of HbF (5)], to one marked by frequent and extremely severe pain crises, acute clinical events, and early mortality. This has led to a number of important studies to gain a better understanding of the natural history of this disease, such as the CSSCD (5, 6). One of the earliest articles from this study revealed the utility of using the frequency of vaso-occlusive crises (termed pain rate) as a surrogate measure for morbidity in SCD, and as an indicator of patients who are likely to have a more severe clinical course (2). This study showed that the pain rate is modulated by a number of factors, including HbF and hematocrit levels (2). Subsequent analyses in the CSSCD also clarified the natural history and identified predictors of ACS, in which patients suffer from hypoxia and chest pain secondary to sickling in the pulmonary vasculature (3), and mortality (4).

In this study, we have examined the effect of five common genetic polymorphisms on HbF level variation in SCD patients. We have confirmed robust genetic associations in two distinct SCD cohorts, totaling >1,500 patients. In the CSSCD, these five SNPs explain >20% of residual HbF variation. Given our current understanding of the distribution of genetic effect sizes for human complex quantitative traits (16–20), this level of phenotypic variation explained by only five genetic polymorphisms is remarkable. The variant with the strongest effect (rs4671393) resides in an intron of the gene *BCL11A*, a gene expressed in erythrocyte precursors and previously implicated in lymphoid malignancies (21, 22). Three of the SNPs (rs28384513, rs9399137, and rs4895441) are located in the intergenic region between *HBS1L* and *MYB* on chromosome 6. Interestingly, two of these SNPs (rs9399137 and rs4895441) are in close proximity (8 kb apart) and were only identified because LD patterns at this genomic region are different between populations of European and African ancestry. This finding is a clear example of the insights that can be gained by using populations of different ethnicities to replicate and refine genetic association signals. For this study, the strategy proved to be instrumental, leading to the identification of an additional SNP (rs4895441) with good predictive power for pain crisis and ACS rates in SCD (Table S2, $P \approx 0.05$). *MYB* has already been well studied for its role in hematopoiesis, whereas *HBS1L* is a poorly characterized gene with unknown biological functions. The fifth SNP is the well known *Xmn*I polymorphism at position −158 in the $^{G}\gamma$-*globin* (*HBG2*) gene promoter (rs7482144).

Understanding the effect that HbF-associated variants have more generally on blood cell production and parameters will be important in defining their biological roles and for the development of targeted therapies. We tested correlations between the five HbF-associated SNPs (Table 1) and several blood cell parameters [RBC count, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), platelet count, monocyte levels]. The only significant associations were between the *BCL11A* rs4671393 and *HBS1L-MYB* rs9399137 SNPs and platelet count in the Brazil cohort and CSSCD, respectively (Table S5). A recent study reported significant associations of *HBS1L-MYB* rs9399137 with RBC indices (RBC count, MCV, MCH), and platelet and monocyte counts in healthy nonanemic Europeans (23). The reason our results contrast with the conclusions from Menzel *et al.* (23) is unclear; it may reflect a difference that exists in how these variants affect nonanemic rather than anemic populations, or may be due to the fact that the authors of the initial study did not correct for the correlation between HbF levels and other RBC-related parameters (23).

In the large prospective CSSCD, we have shown that the high HbF-alleles at five common SNPs from three loci also associate with reduced pain crisis rate in SCD patients. If validated in additional prospective SCD cohorts, this finding will have important implications because, although not an ideal indicator of daily pain in SCD patients (24), pain crisis rate is ostensibly the best predictor of overall morbidity and mortality in SCD (2). Our initial observation that HbF-associated SNPs provide predictive information for pain crises beyond their effect on single measurements of steady-state HbF levels—an effect likely due to the fact that these variants will affect HbF levels over the lifetime of a patient (25, 26)—suggests that clinical genotyping of these variants (and other HbF-associated genetic variants yet to be found) may someday be potentially useful to stratify SCD patients according to severity risk, and to adjust therapeutic strategies accordingly. These variants will need to be further assessed in other large prospective SCD cohorts to fully appreciate their clinical value.

The available sample size of the CSSCD precluded a probing assessment of the effects of these common HbF-associated polymorphisms on ACS and overall survival. It is important that future clinical investigations of the effect of these SNPs on SCD morbidity and mortality be aimed at replicating these findings in large and appropriately powered prospective studies designed for genetic analysis. This suggests that efforts to assemble additional large and well phenotyped SCD cohorts are justified. It will also be fruitful to examine the association of these SNPs with therapeutic responses, particularly to HbF-inducing agents, which show extensive variability in their efficacy. In parallel, basic research to understand the molecular basis by which these genetic variants alter HbF expression is important, as it may lead to better targeted therapies. Finally, our results highlight a potential opportunity to harness the power of genetics to better understand the broad clinical variability in a seemingly "simple" monogenic disease like SCD.

Materials and Methods

Patients. The demographics of the two SCD cohorts analyzed here are given in Table 3. The CSSCD has been described in detail elsewhere (27). HbF, hemoglobin, hematocrit, and white blood cell (WBC) levels were measured at entry and at several subsequent visits. For this study, data from phase 1 of the CSSCD was used. The SCD patients from Brazil were seen at the adult and pediatric SCD clinics at the Hospital de Hematologia da Fundação Hemope, Recife. All of the Brazilian patients were >5 years old, and none were under hydroxyurea treatment when HbF measurements were taken. This project protocol was approved by the Institutional Review Board at the Children's Hospital of Boston.

Phenotypes. To generate the phenotypes used in our analyses for the CSSCD cohort, we averaged all measurements to obtain steady-state levels. We analyzed patients for whom DNA was available: patients with sickle cell anemia (n = 852), sickle cell anemia and α-thalassemia (n = 360), and hemoglobin SC (n = 63) (Table 3, n = 1,275). For HbF, we excluded from the analysis measurements done in patients <5 years old because HbF levels are not yet stable at this early age. To correct for the skewness of the HbF distribution, we $\log_{10}$-transformed and normalized the data to obtain (after correcting for age, gender, and the type of hemoglobinopathy) the quantitative trait used in the association analysis (Fig. S1). For RBC count, platelet count, monocyte levels, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH), only data recorded at entry were analyzed; these hematologic indices were corrected for sex and age, and transformed by using quantile normalization. Pain crisis rate, ACS rate, renal failure, and seizures were determined as previously described (2–4), excluding patient with hemoglobin SC or patients with incomplete genotypic or phenotypic information (Table 3, n = 895). In this set of 895 patients with complete genotype and phenotype information, there were 5,535 pain events, 617 ACS events, and 37 deaths reported. In the Brazil cohort, HbF levels were measured by using HPLC of hemolysates from blood samples. To normalize the percentage of HbF in the Brazil SCD cohort, we square-root transformed the data and corrected for sex and the referring clinic (adult vs. pediatric); the residuals from this correction were used in the genetic association analysis (Fig. S1).

DNA Genotyping. All DNA genotyping was performed by using the mass spectrometry-based MassArray iPLEX platform from Sequenom (28). For SNPs passing quality control, the genotyping success rate was >93% and the consensus error rate, estimated from replicates, was <0.3%. The Hardy–Weinberg equilibrium P value was >0.05 for all SNPs, except for the −158(G>A) *Xmn*I polymorphism in the $^{G}\gamma$-*globin* (*HBG2*) gene promoter (rs7482144), which was in disequilbrium (P = 0.002). This finding is expected in a population of SCD patients for a polymorphism in LD with the sickle cell mutation.

Statistical Analysis. Genetic analysis was performed by using the PLINK software (29), testing only the additive genetic model under a linear regression framework. When several SNPs were genotyped at the same locus, we used conditional analysis and stepwise regression to determine whether one or more independent association signals were present. Statistical models to investigate the relationship between HbF-associated SNPs and SCD complications or survival were built in the statistical package R version 2.5.1. For pain crisis and ACS rates, Poisson regression, controlling for overdispersion, was used and models were compared by means of an analysis of deviance. The difference between the residual deviances of the compared models follows a χ^2 distribution with n degrees of freedom, where n corresponds to the difference in the number of degrees of freedom of the two models (30). To test the effect of genotypes on survival, we used Cox proportional hazards regression (implemented in the R Survival package). The likelihood ratio test was used to assess statistical significance, and models were compared by calculating the −2 log likelihood value of each model. The −2 log likelihood difference between models also follows a χ^2 distribution with n degrees of freedom, where n corresponds to the difference in the number of degrees of freedom of the two models (30).

ACKNOWLEDGMENTS. We thank all of the patients who participated in this study, and we thank members of our laboratories for comments. We thank D. Nathan, D. Altshuler, F. Bunn, and M. Weiss for critical reading of this manuscript; the CSSCD investigators for allowing us access to data and DNA samples from the study; and S. Coady for support. Funding for this project was provided by grants from National Institutes of Health (NIH) and the Howard Hughes Medical Institute (S.H.O.). V.G.S. was supported by a Medical Scientist Training Program award from the NIH.

1. Weatherall DJ, *et al.* (2006) Inherited disorders of hemoglobin. *Disease control priorities in developing countries* (Oxford Univ Press, New York, NY), 2nd Ed, pp 663–680.
2. Platt OS, *et al.* (1991) Pain in sickle cell disease. Rates and risk factors. *N Engl J Med* 325:11–16.
3. Castro O, *et al.* (1994) The acute chest syndrome in sickle cell disease: Incidence and risk factors. The Cooperative Study of Sickle Cell Disease. *Blood* 84:643–649.
4. Platt OS, *et al.* (1994) Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med* 330:1639–1644.
5. Bunn HF (1997) Pathogenesis and treatment of sickle cell disease. *N Engl J Med* 337:762–769.
6. Nathan DG, Orkin SH, Look AT, Ginsburg D (2003) *Nathan and Oski's hematology of infancy and childhood* (Saunders, Philadelphia, PA).
7. Platt OS (2008) Hydroxyurea for the treatment of sickle cell anemia. *N Engl J Med* 358:1362–1369.
8. Pilia G, *et al.* (2006) Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* 2:e132.
9. Thein SL, *et al.* (2007) Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc Natl Acad Sci USA* 104:11346–11351.
10. Menzel S, *et al.* (2007) A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet* 39:1197–1199.
11. Uda M, *et al.* (2008) Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc Natl Acad Sci USA* 105:1620–1625.
12. Labie D, *et al.* (1985) The -158 site 5′ to the G gamma gene and G gamma expression. *Blood* 66:1463–1465.
13. Labie D, *et al.* (1985) Common haplotype dependency of high G gamma-globin gene expression and high Hb F levels in beta-thalassemia and sickle cell anemia patients. *Proc Natl Acad Sci USA* 82:2111–2114.
14. Craig JE, *et al.* (1996) Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nat Genet* 12:58–64.
15. Garner C, *et al.* (1998) Haplotype mapping of a major quantitative-trait locus for fetal hemoglobin production, on chromosome 6q23. *Am J Hum Genet* 62:1468–1474.

16. Willer CJ, *et al.* (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 40:161–169.
17. Kathiresan S, *et al.* (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* 40:189–197.
18. Lettre G, *et al.* (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* 40:584–591.
19. Weedon MN, *et al.* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 40:575–583.
20. Gudbjartsson DF, *et al.* (2008) Many sequence variants affecting diversity of adult human height. *Nat Genet* 40:609–615.
21. Satterwhite E, *et al.* (2001) The BCL11 gene family: Involvement of BCL11A in lymphoid malignancies. *Blood* 98:3413–3420.
22. Liu P, *et al.* (2003) Bcl11a is essential for normal lymphoid development. *Nat Immunol* 4:525–532.
23. Menzel S, *et al.* (2007) The HBS1L-MYB intergenic region on chromosome 6q23.3 influences erythrocyte, platelet, and monocyte counts in humans. *Blood* 110:3624–3626.
24. Smith WR, *et al.* (2008) Daily assessment of pain in adults with sickle cell disease. *Ann Intern Med* 148:94–101.
25. Cohen JC, Boerwinkle E, Mosley TH, Jr., Hobbs HH (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* 354:1264–1272.
26. Kathiresan S, *et al.* (2008) Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med* 358:1240–1249.
27. Farber MD, Koshy M, Kinney TR (1985) Cooperative Study of Sickle Cell Disease: Demographic and socioeconomic characteristics of patients and families with sickle cell disease. *J Chronic Dis* 38:495–505.
28. Campbell CD, *et al.* (2007) Association studies of BMI and type 2 diabetes in the neuropeptide Y pathway: A possible role for NPY2R as a candidate gene for type 2 diabetes in men. *Diabetes* 56:1460–1467.
29. Purcell S, *et al.* (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
30. Faraway JJ (2005) *Linear models in R* (Chapman & Hall/CRC, Boca Raton, FL).