Journal of Human Genetics (2011) 56, 316–323

www.nature.com/jhg

ORIGINAL ARTICLE

Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients

Pallav Bhatnagar[1], Shirley Purvis[2], Emily Barron-Casella[2], Michael R DeBaun[3], James F Casella[2], Dan E Arking[1] and Jeffrey R Keefer[2]

Fetal hemoglobin (HbF) level has emerged as an important prognostic factor in sickle-cell disease (SCD) and can be measured by the proportion of HbF-containing erythrocytes (F-cells). Recently, *BCL11A* (zinc-finger protein) was identified as a regulator of HbF, and the strongest association signals were observed either directly for rs766432 or for correlated single-nucleotide polymorphisms (SNPs). To identify additional independently associated genetic variants, we performed a genome-wide association study (GWAS) on the proportion of F-cells in individuals of African ancestry with SCD from the Silent Infarct Transfusion (SIT) Trial cohort. Our study not only confirms the association of rs766432 (*P*-value $<3.32\times10^{-13}$), but also identifies an independent novel intronic SNP, rs7606173, associated with F-cells (*P*-value $<1.81\times10^{-15}$). The F-cell variances explained independently by these two SNPs are ~13% (rs7606173) and ~11% (rs766432), whereas, together they explain ~16%. Additionally, in men, we identify a novel locus on chromosome 17, glucagon-like peptide-2 receptor (*GLP2R*), associated with F-cell regulation (rs12103880; *P*-value $<3.41\times10^{-8}$). *GLP2R* encodes a G protein-coupled receptor and involved in proliferative and anti-apoptotic cellular responses. These findings highlight the importance of denser genetic screens and suggest further exploration of the *BCL11A* and *GLP2R* loci to gain additional insight into HbF/F-cell regulation.
Journal of Human Genetics (2011) **56,** 316–323; doi:10.1038/jhg.2011.12; published online 17 February 2011

Keywords: *BCL11A*; F-cell regulation; fetal hemoglobin; *GLP2R*; GWAS; sickle-cell disease; SIT Trial cohort

INTRODUCTION

Sickle-cell disease (SCD) is the most common autosomal recessive blood disorder in the United States, affecting approximately 1 in 400 African Americans,[1] and causes considerable morbidity and mortality.[2] The clinical manifestations of SCD include marked phenotypic heterogeneity, with involvement of genetic as well as environmental factors.[3] It is well established that increases in fetal hemoglobin (HbF) can decrease the severity of SCD, because of its ability to inhibit the polymerization of sickle hemoglobin (HbS). Early clinical observations demonstrated that blood from infants with SCD showed little sickling compared with older patients, and that SCD patients with hereditary persistence of HbF production had less severe complications of their disease.[4,5] More recently, Platt *et al.*[2,6] saw a benefit to incremental increases in baseline HbF in terms of painful crises and mortality in patients with SCD. The ameliorating effect of HbF on SCD and other β-hemoglobinopathies has generated intense interest in understanding the control of HbF expression in adults.

HbF predominates in the fetus, but postnatally it declines to extremely low levels and is restricted to a sub-population of erythrocytes termed F-cells.[7] In normal individuals, residual HbF synthesis continues throughout adult life, and HbF and F-cells are closely correlated traits ($r^2>0.9$).[8–10] The heritability of F-cell levels was estimated to be ~90% in an European population,[11] indicating that the expression of the γ-globin gene in adults is under strong genetic control. Previous studies have shown that levels of HbF are influenced by several factors, including age,[12,13] sex[13] and a sequence variant (C→T) at position −158 upstream of the γ-globin gene (11p16), commonly referred as the *Xmn*I-Gγ polymorphism.[14,15] Linkage and association studies for HbF levels mapped quantitative trait loci (QTLs) to chromosome 6q23, 8q and Xp22.2.[16–18] In last few years, several genome-wide association studies (GWAS) in different ethnic groups have established three major QTLs (*BCL11A* at 2p15, *HBS1L-MYB* intergenic region at 6q23 and *Xmn*I-Gγ at 11p16), accounting for 20–50% of phenotypic variation in HbF and F-cell levels.[19–22] Recently, fine mapping of HbF-associated signals confirms the association of these QTLs and ruled out the previously proposed *XmnI*-Gγ polymorphism as an independent causal variant for HbF regulation.[23] Two of these three QTLs, *MYB* and *BCL11A* are oncogenes, and emphasize the importance of cell proliferation and differentiation in HbF/F-cell regulation. Subsequently, *BCL11A* (a zinc-finger transcrip-

[1]McKusick-Nathans Institute of Genetic Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, USA; [2]Department of Pediatrics, Division of Pediatric Hematology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA and [3]Washington University School of Medicine, St Louis, MI, USA
Correspondence: Dr DE Arking, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, School of Medicine, 733 N. Broadway, Room 453, Baltimore, MD 21205, USA.
E-mail: arking@jhmi.edu
or Dr JR Keefer, Division of Pediatric Hematology, Johns Hopkins University, School of Medicine, Ross Building, Room 1129, Baltimore, MD 21205, USA.
E-mail: jrkeefer@jhmi.edu
Received 21 December 2010; accepted 5 January 2011; published online 17 February 2011

tion factor) was shown to function as a regulator of HbF and it was hypothesized that it might repress expression of the γ-globin gene directly by interacting with *cis*-regulatory elements within the β-globin cluster, or indirectly by modulating cellular pathways that affect HbF expression.[24] Interestingly, all the reported significant single-nucleotide polymorphisms (SNPs) of *BCL11A* reside within a region of ~14 kb in intron 2 of the gene, and are in moderate-to-high linkage disequilibrium (LD) in African ancestry in Southwest USA (ASW) samples from the HapMap project,[25] suggesting that they all tag the same genetic signal in that region. To map additional QTLs and gain more insight of the genetic regulation of F-cells, we performed a GWAS on the proportion of F-cells in African ancestry individuals, namely, patients with SCD from the Silent Infarct Transfusion (SIT) Trial cohort.

MATERIALS AND METHODS

Study and population samples

The SIT Trial is an international, multi-center clinical study funded by the National Institute of Neurological Disorders and Stroke (http://sitstudy.wustl.edu/).[26] The study protocol was approved by the Institutional Review Board at the Johns Hopkins University School of Medicine and conducted in accordance with institutional guidelines. Samples were taken pre-transfusion and for each patient, DNA was collected from Epstein–Barr virus transformed lymphoblasts using Puregene Genomic DNA Purification kits (Gentra Systems, Minneapolis, MN, USA). Demographic and phenotypic information were collected for each participant and the inclusion criteria for the recruitment were age (5–15 years) and hemoglobinopathy diagnosis (either Hb SS or Hb $S\beta^0$-thalassemia). Details of the study design are given elsewhere.[26]

Phenotypic assessment of HbF

Peripheral blood drawn from SIT Trial study subjects was mixed 1:1 with Alsever's solution (Sigma, St Louis, MO, USA), stored at 4 °C, and analyzed within 1 week of being drawn. For each subject, F-cells were enumerated using R-Phycoerythrin-conjugated monoclonal antibody directed to HbF (Invitrogen, Camarillo, CA, USA) following the manufacturer's instructions. Negative controls were prepared using isotype-matched nonspecific Phycoerythrin-conjugated antibody (Beckman Coulter, Fullerton, CA, USA). Analysis of 10 000–30 000 cells per tube was performed on a FACScan flow cytometer (Becton Dickinson, Franklin Lakes, NJ, USA) and data was analyzed using CellQuest software (Becton Dickinson). Fetaltrol (Trillium diagnostics, LLC, Brewer, ME, USA) was used as a tri-level fetal red cell control following the manufacturer's instructions. All of the F-cell determinations were carried out in a central laboratory at the time of the SIT Trial blood draw.

Genotyping

Samples were genotyped with the Illumina HumanHap650Y array (Illumina Inc., San Diego, CA, USA), which interrogates approximately 661 000 SNPs, of which ~100 000 were selected as tags for populations with African ancestry.[27] The Beadstudio software (Illumina Inc.) was used to cluster the data, and samples with <96.5% call rates were re-genotyped. A total of 24 International HapMap Consortium[25] controls and 13 known duplicates were also genotyped. The reproducibility, calculated from duplicate pairs, was 99.98% and genotype concordance with HapMap data was 99.76%.

Quality control

Cryptic relatedness in the cohort was determined by examining pairwise identity-by-state, and 33 samples were identified as first-degree relatives (full or half siblings) and excluded from the study. Additionally, eight samples, because of missing covariate data, were also excluded from the analysis. Given the admixed nature of African Americans, we used principal component analysis as implemented in EIGENSTRAT[28] to both identify genetic outliers (>6 s.d.'s on any of the top 10 principal components) and correct for any potential residual population substructure. Six individuals were identified as genetic outliers and were dropped from the study, leaving 440 individuals for the subsequent analysis. Variance inflation factor for genomic control (λ_{GC}) was estimated as described by Devlin and Roeder[29] to test for residual relatedness and/or population substructure.

Statistical analysis

All association analysis and quality control measures were performed using the PLINK software package,[30] version 1.06 (http://pngu.mgh.harvard.edu/purcell/plink/). F-cell distribution in the studied samples was slightly skewed; therefore, Box–Cox power transformation (λ=0.6) was applied to approximate the normal distribution.[31] The effect of each SNP on the proportion of F-cell levels was assessed by adjusting for age, sex and top 10 principal components in a multivariate linear regression by assuming an additive genetic model of inheritance. The genome-wide significant threshold was determined by permutation (P-value $<1.27\times10^{-7}$). R statistical computing environment (http://www.r-project.org/) (version 2.9.0) was used to generate quantile–quantile (Q–Q), Manhattan and regional association plots. To create a more comprehensive fine map of the SNPs from the observed genome-wide significant loci, imputation was performed by using Hidden Markov model as implemented in the MACH software (version 1.0.16) (http://www.sph.umich.edu/csg/abecasis/MACH/).[32] The YRI and CEU combined panel, from the 1000 Genomes Project,[33] was used as a reference population, and 100 iterations were used to estimate model parameters. To account for the uncertainty of imputed data, the estimated allele dosage was analyzed using ProABEL[34] under a linear regression framework. Standard quality metrics were applied and only SNPs with high-quality ($r^2>0.8$) score were considered for the analysis. Further, the LD patterns within the surrounding region of the significant SNPs were constructed using the solid spine method,[35] as implemented in Haploview[36] (version 4.1) (http://www.broad.mit.edu/mpg/haploview/index.php). Haplotype inferences were carried out using a Bayesian statistical method implemented in PHASE software (version 2.1) (http://www.stat.washington.edu/stephens/).[37] Default settings of 100 iterations, 100 burn-in steps and 1 thinning interval were used to infer most likely pairs of haplotypes for each individual. Inferred haplotype diversity was represented by means of a cladogram, which is constructed using hierarchical Ward's clustering method. Haplotype-based association analysis was performed using generalized linear regression, assuming an additive genetic model, and adjusted for age, sex and the first 10 principal components.

RESULTS

Genome-wide single SNP association

Genome-wide association was performed for the proportion of F-cells in 440 individuals (232 men, 208 women) from the SIT Trial cohort. The average age of the cohort was 9.15 years, with 53% males. Detailed demographic and clinical characteristics for the study subjects are described in Table 1. A Q–Q plot of the observed P-values versus expected is shown in Figure 1a. The observed P-values show no early departure from the null, suggesting that our findings are unlikely to be influenced by poor genotyping, sample relatedness or population stratification. The genomic control (λ_{GC}=1.001) from the

Table 1 Demographic and clinical characteristics

Traits	*SIT Trial cohort*[a]
Sex, *n* (%)	
Men	232 (52.7)
Women	208 (47.3)
Age (in years), mean ± s.d.	9.15 ± 2.1
Sickle hemoglobinopathy, n (%)	
Hb SS	382 (86.8)
Hb $S\beta^0$	32 (7.2)
Unknown	26 (6)
Hematocrit (%), mean ± s.d.	23.2 ± 3.3
Hemoglobin ($g\,dl^{-1}$), mean ± s.d.	8.0 ± 1.0
Fetal hemoglobin (%), mean ± s.d.	8.57 ± 5.5
F-cells (%), mean ± s.d.	41.1 ± 19.6
F reticulocytes (%), mean ± s.d.	17.9 ± 14.5
White blood cell counts (Cu)	14020 ± 13712
Reticulocytes (%), mean ± s.d.	11.7 ± 5.2

Abbreviations: Hb SS, sickle-cell anemia; Hb $S\beta^0$, sickle-beta plus thalassaemia; SIT, Silent Infarct Transfusion.
[a]Sample exclusion criteria for the SIT Trial cohort: non-African–American ancestry, all thalassemias, except Hb SS or Hb $S\beta^0$, first-degree relatives and age younger than 5 years and older than 15 years.

analyzed 660 740 SNPs also suggests minimal population stratification. The distribution of association *P*-values (Manhattan plot) for F-cell levels is shown in Figure 1b. We observe a genome-wide significant finding at the previously implicated chromosome 2p15 region and confirm the association of the *BCL11A* locus in the modulation of F-cell levels. In total, four SNPs from this locus were observed to be genome-wide significant (permuted threshold $<1.27\times10^{-7}$), and includes the SNP, rs766432 (*P*-value $<3.32\times10^{-13}$), which previously has been reported to be associated with HbF/F-cell levels in diverse populations.[19,22,38,39] We also identified an additional *BCL11A* intronic SNP (rs6706648), with more significant association with F-cell levels (*P*-value $<4.71\times10^{-14}$) (Table 2). In our study, the ancestral alleles of both rs766432 and rs6706648 SNPs are observed to be associated with lower F-cell levels (rs766432, $\beta=-1.49$; rs6706648, $\beta=-1.42$) (Table 2). Distributions of the proportion of F-cell levels within each genotype group of these two SNPs are shown in Figure 2a. Individuals who were homozygous for ancestral alleles at rs6706648 (TT) and rs766432 (AA) loci have ~two times lower F-cells, when compared to those who were homozygous for the derived alleles at both loci (Figure 2b). In our samples, we did not observe any individual carrying homozygous-derived alleles at rs766432 and homozygous ancestral alleles (TT) at the rs6706648 locus. Our sentinel SNP, rs6706648, is located in intron 2 of the *BCL11A* region and

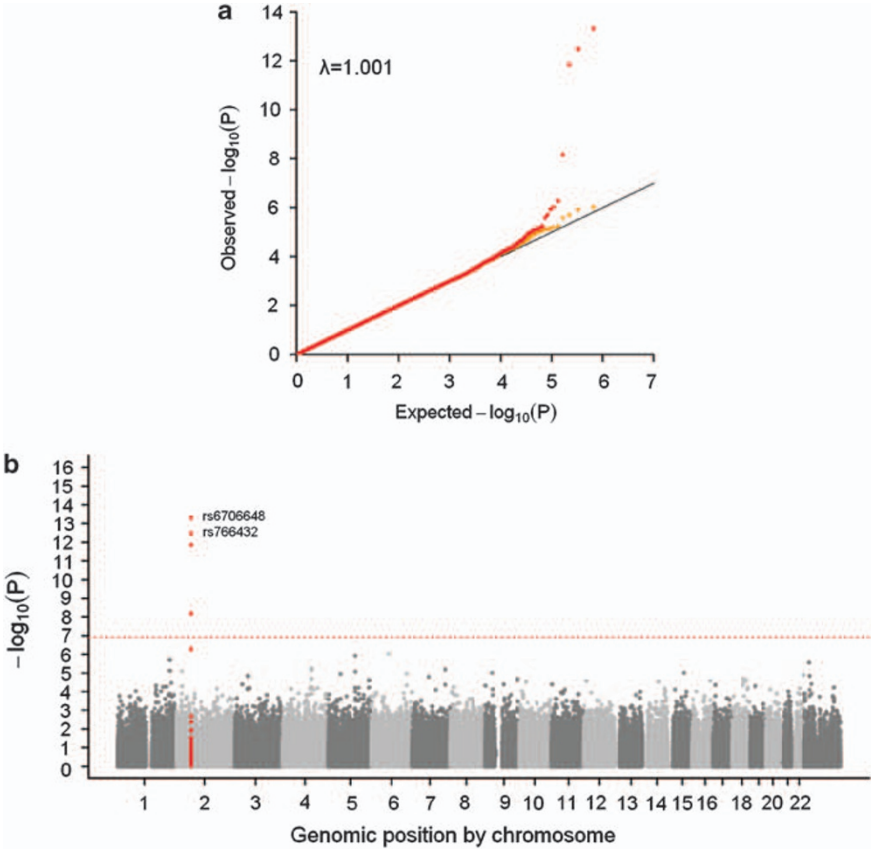
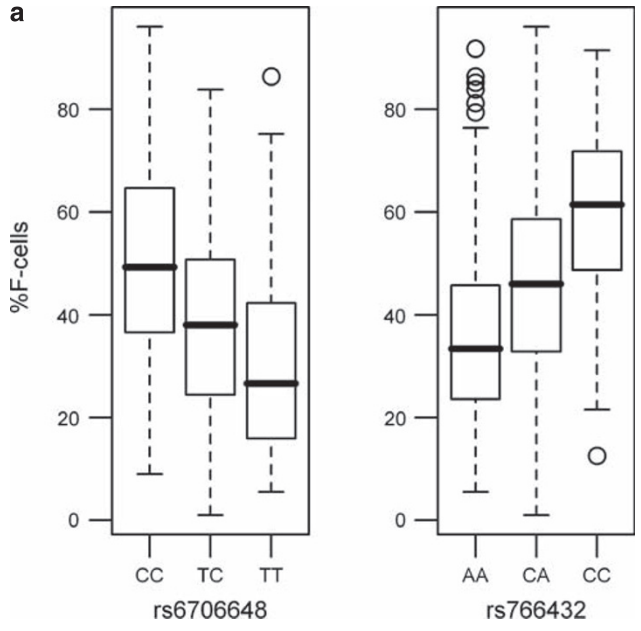


Figure 1 Summary of the genome-wide association results of the proportion of F-cells in the SIT Trial cohort. (**a**) Q–Q plot of the observed versus the expected *P*-values from an additive genetic model for the entire set of 660 740 SNPs (red), and after removing genome-wide significant and their neighboring ± 100 kb region SNPs (yellow). (**b**) Manhattan plot for F-cells association results plotted against the position on each chromosome. The red color peak on chromosome 2 corresponds to the *BCL11A* region (± 100 kb SNPs from rs6706648) and the red horizontal line represents a permutation-based genome-wide significant threshold (*P*-value $<1.27\times10^{-7}$).

Table 2 Genome-wide significant SNPs associated with the proportion of F-cells in the SIT Trial cohort

Chr.	*SNP*	*Position*[a]	*Gene*	*Location*	*Coded allele (ancestral)*	*Non-coded allele (derived)*	*Coded allele frequency*	β-*value*[b]	*s.e.*	P-*value*
2	rs766432	60 573 474	*BCL11A*	Intron	A	C	0.74	−1.49	0.20	3.32×10^{-13}
2	rs10195871	60 574 093	*BCL11A*	Intron	G	A	0.71	−1.42	0.20	1.40×10^{-12}
2	rs6706648	60 575 544	*BCL11A*	Intron	T	C	0.41	−1.42	0.18	4.71×10^{-14}
2	rs6709302	60 581 133	*BCL11A*	Intron	G	A	0.66	1.15	0.20	6.69×10^{-9}

Abbreviations: Chr.,chromosome; SNP, single-nucleotide polymorphism; SIT, Silent Infarct Transfusion.
[a]Positions are in reference to UCSC human genome hg18 coordinates.
[b]Effect size of the coded allele based on multivariate linear regression adjusted for age, sex and first 10 principal components.



b

rs6706648 \ rs766432	AA	CA	CC
CC	42.8 (50)	52.3 (63)	59.3 (30)
TC	35.9 (119)	41.4 (104)	-
TT	31.1 (74)	-	-

Figure 2 Distribution of F-cells by observed genotypes. (**a**) Box plot demonstrating the distribution of untransformed F-cells within each genotype group of rs6706648 and rs766432 SNPs. Each box represents the F-cell values between the 25th and 75th quartiles, and the dark black line within the boxes indicates the median values. (**b**) Mean distribution of the untransformed F-cells in individuals by genotype combination. The number of individuals within each genotype combination is shown in parentheses.

observed in low LD with rs766432 (r^2=0.25) (Figure 3a). The additional two significantly associated SNPs from the same region, rs6709302 and rs10195871, were observed in moderate-to-high LD with either rs6706648 (r^2=0.52) or rs766432 (r^2=0.85), respectively; hence, likely tagging the same signal (Supplementary Figure 1). To test for independent genetic effects of rs6706648 and rs766432 in the *BCL11A* region, we performed conditional multivariate regression analysis. The conditional analyses results are shown in Table 3, and indicate that both SNPs (rs6706648 and rs766432) have independent genetic effects. Independently, these two SNPs explain ~12% (rs6706648) and ~11% (rs766432) of the F-cell variance in patients with SCD, whereas together they explain ~15%. Although no other loci are genome-wide significant, several loci showed suggestive association with F-cell levels (Supplementary Table 1).

Imputation-based association testing

Using the YRI and CEU combined reference panel from the 1000 genomes project, imputation was performed over a 90-kb interval centered on rs6706648. In total, 102 SNPs were imputed from the *BCL11A* region and after dropping low-quality imputed SNPs ($r^2<0.8$), 42 SNPs were analyzed. Among these high-quality imputed SNPs, 17 common variants were observed significant at the genome-wide threshold (Supplementary Table 2). The strongest evidence for association was observed for rs7606173 (*P*-value $<5.14\times10^{-16}$), which is ~3.4 kb downstream to rs6706648 and is in high correlation (r^2=0.91) (Figure 3b, Supplementary Figure 1). In our study, we replicate the association of all previously reported *BCL11A* SNPs[19–23] and no other SNP (except rs7606173) was observed to be more significant than rs6706648 (Supplementary Table 2). The ancestral allele (G) of rs7606173 is the major allele (frequency: 0.55) and is associated with higher F-cell levels (β=1.50) (Supplementary Table 2). Among 17 genome-wide significant SNPs, 10 SNPs were observed with high LD ($r^2>0.85$) to rs766432 (Supplementary Table 2, Supplementary Figure 1). To identify independent genetic effects, conditional regression analysis was performed on imputed and genotyped SNPs. Using conditional regression, we demonstrate that rs7606173 and rs766432 account for the genetic effects on F-cell levels in the *BCL11A* region (Supplementary Table 3). The variance explained by rs7606173 is ~13%, and together with rs766432 they explain ~16% of F-cell variability.

Haplotype association analysis

To gain additional insight into the genetic association observed in the *BCL11A* intron 2 region, haplotype-based association analysis was performed. In total, seven haplotypes (frequency >1%) were inferred from the LD block containing rs766432 and rs7606173 (Figure 4). To generate haplotype clusters, the ancestral allele for each SNP was determined unambiguously by comparing sequence similarity with non-human primates. On this basis, all the inferred haplotypes are grouped into three clusters (I, II and III). Cluster I contains ~42% of total haplotypes and has ancestral alleles at the majority of the loci, and was therefore used as the reference haplotype cluster. Using a linear regression framework, we observed that haplotypes from the reference cluster are associated with the lowest F-cell levels, whereas cluster III haplotypes were observed with the highest levels (Figure 4). As expected, cluster II haplotypes, containing derived alleles at both

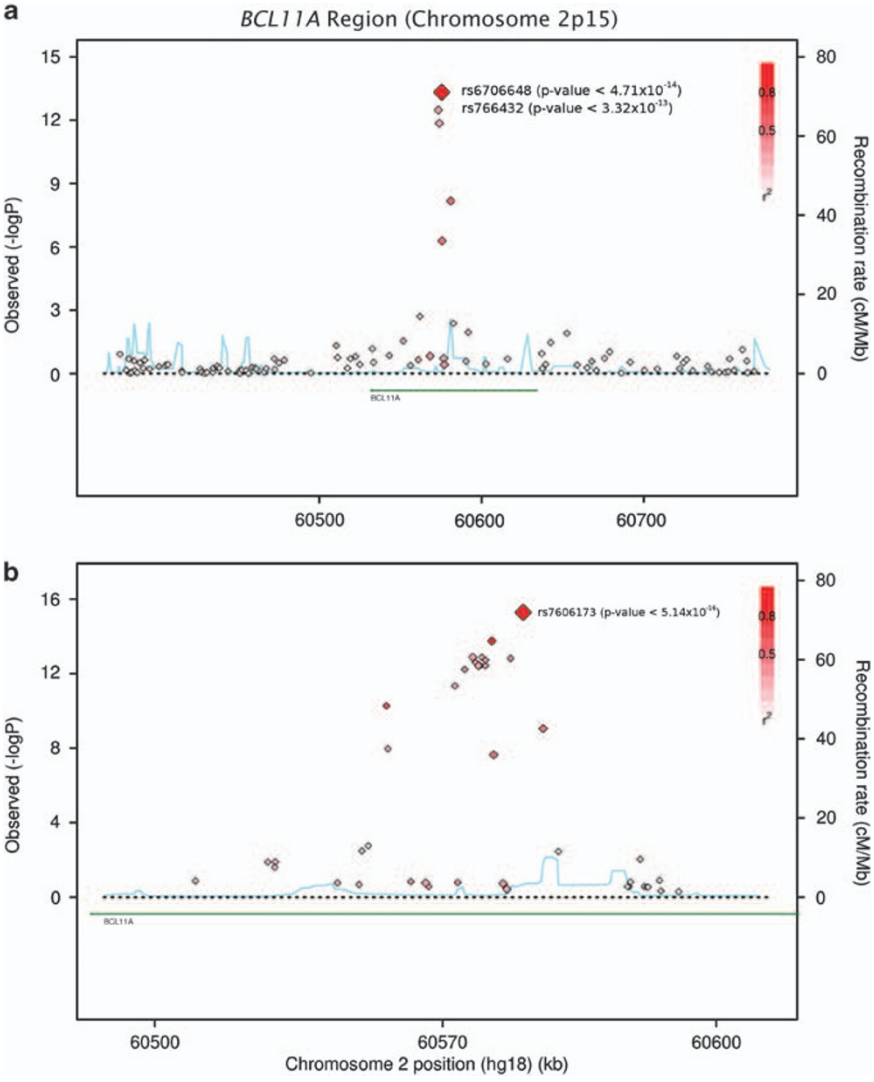


Figure 3 Regional association plot of genotyped and imputed SNPs from the *BCL11A* intron 2 region. (**a**) Genotyped SNPs are plotted with their *P*-values ($-\log_{10} P$) as a function of genomic position (UCSC human genome hg18 coordinates). Estimated recombination rates observed in HapMap YRI (using a window of ± 200 Kb) are plotted to reflect the local LD structure around rs6706648. (**b**) Regional association plot showing the significance of the imputed *BCL11A* SNPs. The significance of the imputed SNPs is plotted with their *P*-values ($-\log_{10}$ values) as a function of genomic position (UCSC human genome hg18 coordinates). Estimated recombination rates observed in HapMap YRI are plotted to reflect the local LD structure around the most significant SNP (rs7606173) observed in our study. The imputation of the plotted SNPs were performed using the YRI and CEU combined reference panel from the 1000 Genomes Project.[33]

Table 3 Association summary of genome-wide significant SNPs from conditional multivariate regression analysis

						Conditional analysis[a]					
						rs6706648			*rs766432*		
Chr.	*SNP*	*Position*[b]	*Coded allele (ancestral)*	*Non-coded allele (derived)*	*Pairwise LD (r^2)*[c]	*β-value*	*s.e.*	*P-value*	*β-value*	*s.e.*	*P-value*
2	rs766432	60 573 474	A	C	0.25	−0.99	0.22	7.60×10^{-6}	—	—	—
2	rs10195871	60 574 093	G	A	0.28	−0.9	0.22	4.58×10^{-5}	−0.5	0.51	0.3257
2	rs6706648	60 575 544	T	C	—	—	—	—	−0.99	0.2	1.04×10^{-6}
2	rs6709302	60 581 133	G	A	0.51	0.29	0.26	0.2536	0.78	0.2	8.26×10^{-5}

Abbreviations: Chr.,chromosome; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism; SIT, Silent Infarct Transfusion.
[a]Conditional analysis was performed adjusting for age, sex and first 10 principal components.
[b]Positions are in reference to UCSC human genome hg18 coordinates.
[c]Pairwise LD of each SNP against rs6706648 (from 440 SIT Trial individuals).

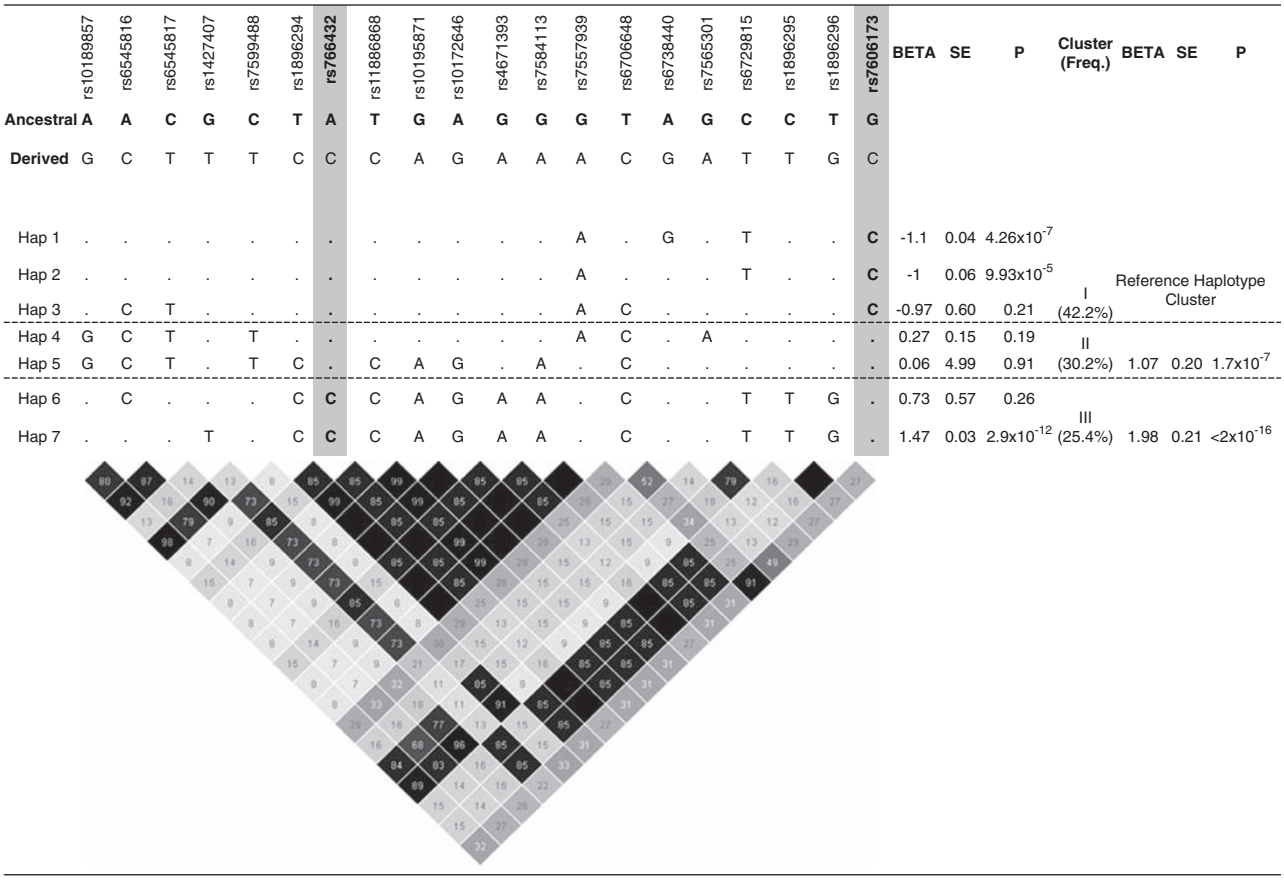


Figure 4 Genetic association of inferred haplotypes from the rs766432 and rs7606173 region LD block. In total, seven common haplotypes (frequency >1%) were inferred using 20 SNPs from the rs766432 and rs7606173 region LD block. The top panel provides the information of the ancestral alleles based on the similarity with non-human primates. The period symbol ('.') indicates the ancestral allele and letters in place of period symbol indicate derived alleles. The inferred haplotypes are grouped in three clusters (I, II and III) and cluster I is defined as the 'reference haplotype cluster'.

rs766432 and rs7606173 loci, were observed with an intermediate effect. Further, in reference to the ancestral cluster (cluster I), we observed an apparent additive effect for clusters II and III. The effect size estimate of cluster III haplotypes is ~two-fold higher (β=1.98) than that observed for cluster II haplotypes (β=1.07) (Figure 4). The F-cell variances explained by cluster III haplotypes alone and together with cluster II haplotypes were ~11 and ~16%, respectively.

Sex-stratified analysis

Given known sex-specific differences in HbF/F-cell levels,[13] we also performed sex-stratified genome-wide analysis (Supplementary Tables 4 and 5). In men, we report an additional locus on chromosome 17p13.3, glucagon-like peptide-2 receptor (*GLP2R*), showing genome-wide significance for F-cell levels (rs12103880; *P*-value $<3.41\times10^{-8}$) (Supplementary Table 5). SNP rs12103880 is located at the 5′ UTR region of *GLP2R* and the ancestral allele (G) is associated with lower F-cell levels (β=−1.36) (Supplementary Table 5).

Association with previously reported loci

To date, several SNPs from three major QTLs (2p15, 6q23 and 11p16) have been reported in association with HbF/F-cells,[19–22] though many of these SNPs are in moderate-to-high LD and are likely tagging the same genetic signal at each locus. We estimated the correlations among reported SNPs based on ASW HapMap (phase II and III) data, and an association summary of these SNPs ($r^2<0.3$) is shown in Supplementary Table 6. From the 2p15 region, other than independent genome-wide significant rs7606173 and rs766432 SNPs, we also observed an association with the same direction effect for SNP rs6732518 (*P*-value $<9.5\times10^{-4}$) (Supplementary Table 6). Similarly, significant association from the 6q23 region was observed for rs9399137 (*P*-value <0.001) and rs4895441 (*P*-value $<5.0\times10^{-4}$), even after Bonferroni correction for the number of independent loci tested (α=0.05/13, *P*-value$\leqslant$0.0038), confirming the role of genetic variation in *MYB-HBS1L* in the regulation of F-cell levels.

DISCUSSION

In recent years, remarkable progress has been made through the use of genome-wide scans, demonstrating the feasibility of an unbiased approach to identify novel targets for therapeutic interventions.[40,41] Given these successes, several GWAS have been attempted to elucidate the genetic regulation of HbF and F-cell levels in diverse populations.[19,20,22,23] Here, we present results of a GWAS conducted on 440 individuals from the SIT Trial cohort. This study represents the largest genome-wide scan of the proportion of F-cell levels in patients with SCD of African ancestry.

Our results confirm the robust genetic association of *BCL11A* in the modulation of F-cell levels. A strong effect of this region was originally seen in non-anemic Europeans[19] and subsequently replicated in other populations.[20–22] Through this study, we not only confirm the previously associated *BCL11A* SNP (rs766432), but also report an independent effect of rs7606173 in F-cell level regulation (Supplementary Table 2). The observed effect of rs766432 is in the same

direction as reported in previous GWAS and candidate gene studies performed in African Americans.[22,38] Recently, as part of a HbF regulation fine mapping association study in an African American cohort with SCD, Galarneau *et al.*[23] reported the strongest genetic effect in *BCL11A* at intronic SNP rs4671393. Using stepwise conditional regression, they also reported two additional SNPs (rs7599488 and rs10189857) exhibiting independent genetic effects on HbF levels. In our study, we observed similar association effects for these SNPs, and rs4671393 is in perfect LD with rs766432 ($r^2=1$) (Supplementary Tables 2 and 3, Supplementary Figure 1). All three SNPs reported by Galarneau *et al.*[23] are in the same LD block as our two independent SNPs (rs766432 and rs7606173) (Figure 4). It is noteworthy that, after conditioning on the genetic effect of rs766432, we observed the enhanced genetic significance of rs7599488 and rs10189857 observed by Galarneau *et al.*[23], but this significance does not hold when the additional effect of rs7606173 is included (Supplementary Table 3). Our result suggests rs7606173 captures the entire genetic effects of these reported SNPs, and rs766432 and rs7606173 represent two independent effects. This locus demand further functional validation for a better understanding of *BCL11A* in the regulation of HbF levels.

Given the previous evidence of sex-specific differences in HbF/F-cell levels,[13] we performed sex-stratified analysis and report a novel locus from the chromosome 17p13.3 (*GLP2R*) region, associated with F-cell levels in men (rs12103880; *P*-value $<3.41\times10^{-8}$) (Supplementary Table 5). *GLP2R* encodes a G protein-coupled receptor that participates in cellular signaling through multiple G proteins to affect the cyclic adenosine monophosphate and mitogen-activated protein kinase pathways, leading to both proliferative and anti-apoptotic cellular responses. Accounting the cell proliferative and oncogenic function of the two major previously reported QTLs (*MYB* and *BCL11A*), the genetic association of *GLP2R* region seems to be of high relevance and needs to be replicated in larger samples.

Other than *BCL11A*, we did not observe any genome-wide significant association of previously reported QTLs. Though our study failed to identify genome-wide significant SNPs (permuted threshold $<1.27\times10^{-7}$) at chromosome 6q23 (*MYB-HBS1L* intergenic region), under a candidate SNP approach, signals from rs9399137 (*P*-value <0.001) and rs4895441 (*P*-value $<5.0\times10^{-4}$) loci support their genetic involvement in F-cell regulation (Supplementary Table 6). Additionally, on chromosome X, we observed moderately strong association for rs12559632 (*P*-value $<2.64\times10^{-6}$) and rs6630120 (*P*-value $<1.48\times10^{-5}$) SNPs from *PHEX* (Xp22.2) and *MAGEB18* (Xp21.3) genes, respectively (Supplementary Table 1). Our observation for the potential involvement of *PHEX* region in F-cell modulation is in agreement with a recent GWAS performed in an African American cohort.[22] Interestingly, both these genes are close to the Xp22.2 locus, which was identified by a linkage study and associated with the regulation of HbF levels.[18]

This study differs from recent GWAS reports in patients with SCD in that F-cell number rather than total HbF was used exclusively as the HbF phenotype. The F-cell phenotype was used by Menzel *et al.*[19] in the initial identification of *BCL11A* in a non-SCD population. Total HbF levels in patients with SCD are a reflection of three independently regulated factors: F-cell production rate, preferential survival of F-cells and the amount of HbF per F-cell.[42] Results from GWAS studies, using the total HbF phenotype, might identify genes contributing to any of these processes, whereas our study identifies genes contributing to F-cell production or preferential F-cell survival, but not HbF per F-cell. We believe that focusing on F-cells is a rational approach. The high correlation between F-cell number and total HbF ($r^2>0.9$)[8–10] suggests that HbF variation is determined largely by F-cell number, not HbF per F-cell. Additionally, given that the goal of HbF-based therapy for SCD is to reduce polymerization of HbS in as many cells as possible, manipulation of F-cell number rather than the amount of HbF per F-cell is likely to have much greater impact on the course of SCD. Patients with SCD already have relatively high HbF per F-cell (average of 38%),[42] which should be sufficient to inhibit HbS polymerization in F-cells.

In summary, we report a novel independent genetic variant (rs7606173) associated with F-cell regulation, and, in men, a novel locus at 17p13.3 (*GLP2R*) is associated with F-cell levels. Additionally, we validate the genetic significance of *BCL11A* region (2p15) and *MYB-HBS1L* SNPs. This study highlights the importance of denser genetic screens of the *BCL11A* region in large and well-powered studies. Confirmation of these variants might help to improve the prediction of one's ability to produce HbF in response to disease, and will have implications for prenatal diagnosis and genetic counseling of patients with SCD.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank the staff, clinicians and patients for their participation in the Silent Infarct Transfusion (SIT) Trial study. This study was supported by the National Heart, Lung and Blood Institute (NHLBI) (award number: U54HL090515, 5R01HL091759) and the National Institute of Neurological Disorders and Stroke (NINDS) (NIH-NINDS 5U01-NS042804-03). Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University (CIDR contract number: HHSN268200782096C).

1 Motulsky, A. G. Frequency of sickling disorders in U.S. blacks. *N. Engl. J. Med.* **288,** 31–33 (1973).
2 Platt, O. S., Brambilla, D. J., Rosse, W. F., Milner, P. F., Castro, O., Steinberg, M. H. *et al.* Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N. Engl. J. Med.* **330,** 1639–1644 (1994).
3 Driss, A., Asare, K. O., Hibbert, J. M., Gee, B. E., Adamkiewicz, T. V. & Stiles, J. K. Sickle cell disease in the post genomic era: a monogenic disease with a polygenic phenotype. *Genomics Insights* **30,** 23–48 (2009).
4 Watson, J. The significance of the paucity of sickle cells in newborn Negro infants. *Am. J. Med. Sci.* **215,** 419–423 (1948).
5 Jacob, G. F. & Raper, A. B. Hereditary persistence of foetal haemoglobin production, and its interaction with the sickle-cell trait. *Br. J. Haematol.* **4,** 138–149 (1958).
6 Platt, O. S., Thorington, B. D., Brambilla, D. J., Milner, P. F., Rosse, W. F., Vichinsky, E. *et al.* Pain in sickle cell disease. Rates and risk factors. *N. Engl. J. Med.* **325,** 11–16 (1991).
7 Boyer, S. H., Belding, T. K., Margolet, L. & Noyes, A. N. Fetal haemoglobin restriction to a few erythrocytes (F cells) in normal human adults. *Science* **188,** 361–363 (1975).
8 Miyoshi, K., Kaneto, Y., Kawai, H., Ohchi, H., Niki, S., Hasegawa, 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).
9 Zago, M. A., Wood, W. G., Clegg, J . B., Weatherall, D. J., O'Sullivan, M. & Gunson, H. Genetic control of F cells in human adults. *Blood* **53,** 977–986 (1979).
10 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).
11 Garner, C., Tatu, T. & Reittie, J. E. Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood* **95,** 342–346 (2000).
12 Rutland, P. C., Pembrey, M. E. & Davies, T. The estimation of fetal haemoglobin in healthy adults by radioimmunoassay. *Br. J. Haematol.* **53,** 673–682 (1983).
13 el-Hazmi, M. A., Warsy, A. S., Addar, M. H. & Babae, Z. Fetal haemoglobin level-effect of gender, age and haemoglobin disorders. *Mol. Cell. Biochem.* **135,** 181–186 (1994).
14 Gilman, J. G. & Huisman, T. H. J. DNA sequence variation associated with elevated fetal $G\gamma$ globin production. *Blood* **66,** 783–787 (1985).
15 Garner, C., Tatu, T., Game, L., Cardon, L. R., Spector, T. D., Farrall, M. *et al.* A candidate gene study of F cell levels in sibling pairs using a joint linkage and association analysis. *Gene Screen* **1,** 9–14 (2000).

16 Craig, J. E., Rochette, J., Fisher, C. A., Weatherall, D. J., Marc, S., Lathrop, M. *et al.* Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nat. Genet.* **12,** 58–64 (1996).
17 Garner, C. S., Menzel, C., Martin, C., Silver, N., Best, S., Spector, T. D. *et al.* Interaction between two quantitative trait loci affects fetal haemoglobin expression. *Ann. Hum. Genet.* **69,** 707–714 (2005).
18 Dover, G. J., Smith, K. D., Chang, Y. C., Purvis, S., Mays, A., Meyers, D. A. *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,** 816–824 (1992).
19 Menzel, S., Garner, C., Gut, I., Matsuda, F., Yamaguchi, M., Heath, 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).
20 Uda, M., Galanello, R., Sanna, S., Lettre, G., Sankaran, V. G., Chen, W. *et al.* Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia.. *Proc. Natl Acad. Sci.* **105,** 1620–1625 (2008).
21 Lettre, G., Sankaran, V. G., Bezerra, M . A., Araujo, A. S., Uda, M., Sanna, S. *et al.* DNA polymorphisms at the BCL11A, HBS1L-MYB, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl Acad. Sci.* **105,** 11869–11874 (2008).
22 Solovieff, N., Milton, J. N., Hartley, S. W., Sherva, R., Sebastiani, P., Dworkis, D. A. *et al.* Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5′ olfactory receptor gene cluster. *Blood* **115,** 1815–1822 (2010).
23 Galarneau, G., Palmer, C. D., Sankaran, V. G., Orkin, S. H., Hirschhorn, J. N. & Lettre, G. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42,** 1049–1051 (2010).
24 Sankaran, V. G., Menne, T. F., Xu, J., Akie, T. E., Lettre, G., Handel, B. V. *et al.* Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science* **322,** 1839–1842 (2008).
25 International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
26 Casella, J. F., King, A. A., Barton, B., White, D. A., Noetzel, M. J., Ichord, R. N. *et al.* Design of the silent cerebral infarct transfusion (SIT) trial. *Pediatr. Hematol. Oncol.* **27,** 69–89 (2010).
27 Steemers, F. J. & Gunderson, K. L. Whole genome genotyping technologies on the BeadArray platform. *Biotechnol. J.* **2,** 41–49 (2007).
28 Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
29 Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
30 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
31 Box, G.E.P. & Cox, D. R. An analysis of transformations. *J. R. Stat. Soc. Series B.* **26,** 211–252 (1964).
32 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).
33 Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
34 Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C. M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11,** 134 (2010).
35 Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).
36 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).
37 Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68,** 978–989 (2001).
38 Sedgewick, A. E., Timofeev, N., Sebastiani, P., So, J. C. C., Ma, E.S.K., Chan, L. C. *et al.* BCL11A is a major HbF quantitative trait locus in three different populations with beta-hemoglobinopathies. *Blood Cells Mol. Dis.* **41,** 255–258 (2008).
39 Nuinoon, M., Makarasara, W., Mushiroda, T., Setianingsih, I., Wahidiyat, P., Sripichai, O. *et al.* A genome-wide association identified the common genetic variants influence disease severity in β^0-thalassemia/hemoglobin E. *Hum. Genet.* **127,** 303–314 (2010).
40 Hardy, J. & Singleton, A. Genomewide association studies and human disease. *N. Engl. J. Med.* **360,** 1759–1768 (2009).
41 Arking, D. E. & Chakravarti, A. Understanding cardiovascular disease through the lens of genome-wide association studies. *Trends Genet.* **25,** 387–394 (2009).
42 Dover, G. J., Boyer, S. H., Charache, S. & Heintzelman, K. Individual variation in the production and survival of F cells in sickle-cell disease. *N. Engl. J. Med.* **299,** 1428–1435 (1978).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)