nature genetics

Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation

Geneviève Galarneau[1], Cameron D Palmer[2,3], Vijay G Sankaran[4,5], Stuart H Orkin[4–6], Joel N Hirschhorn[2,3,7] & Guillaume Lettre[1,8]

We used resequencing and genotyping in African Americans with sickle cell anemia (SCA) to characterize associations with fetal hemoglobin (HbF) levels at the *BCL11A, HBS1L-MYB* and β-globin loci. Fine-mapping of HbF association signals at these loci confirmed seven SNPs with independent effects and increased the explained heritable variation in HbF levels from 38.6% to 49.5%. We also identified rare missense variants that causally implicate *MYB* in HbF production.

HbF is a strong and heritable modifier of disease severity for individuals with sickle cell disease (SCD, including sickle cell anemia (HbSS) but also HbSC and HbS-β-thalassemia) and β-thalassemia; individuals with high HbF levels have less severe complications and a longer life expectancy. Three loci (at *BCL11A*, *HBS1L-MYB* and β-globin) carry DNA polymorphisms that modulate HbF levels[1–4]. To fine map the HbF association signals, we resequenced 175.2 kb from these loci in 190 individuals, including the HapMap European CEU and Nigerian YRI founders and 70 African Americans with SCA (**Supplementary Methods**). We discovered 1,489 DNA sequence variants, including 910 previously unreported variants (**Supplementary Fig. 1** and **Supplementary Tables 1** and **2**). Using this information and data from HapMap, we selected and genotyped 95 SNPs in 1,032 African Americans with SCA (**Supplementary Methods**). We genotyped 17 and 35 SNPs at the *BCL11A* and *HBS1L-MYB* loci, respectively, to characterize previously reported HbF association signals[4]. We also genotyped 43 SNPs at the β-globin locus to capture the majority of the common genetic variation on the main sickle cell haplotypes. Association results are presented in **Supplementary Table 3**.

BCL11A is a direct repressor of HbF production[5] and a major regulator of developmental globin gene switching[6]. Consistent with previous reports[3,4], rs4671393 in *BCL11A* intron 2 was the genetic marker most strongly associated with HbF levels ($P = 3.7 \times 10^{-37}$) (**Table 1**). Stepwise conditional analyses found two other SNPs (rs7599488 and rs10189857) which independently associated with HbF levels (**Table 1**). These two SNPs, located in *BCL11A* intron 2, are in weak linkage disequilibrium (LD) with rs4671393 ($r^2 = 0.17$ and $r^2 = 0.15$ for rs7599488 and rs10189857, respectively) but are in strong LD with each other ($r^2 = 0.96$). When we used principal component analysis to control for admixture, we observed only minor differences in the results (**Supplementary Table 4**).

To further understand the contribution of rs10189857, rs7599488 and rs4671393 to the *BCL11A* HbF association signal, we performed a haplotype analysis. These three SNPs form four haplotypes that represent 99.7% of all haplotypes at this locus. These haplotypes were more strongly associated with HbF levels ($P = 4.0 \times 10^{-45}$) than rs4671393 ($P = 3.7 \times 10^{-37}$) and explained 18.1% of the phenotypic variation in HbF levels (**Supplementary Table 5**). Thus, these haplotypes explain more phenotypic variance than the cumulative sum of the three *BCL11A* SNPs taken individually (14.7%) (**Table 1** and **Supplementary Methods**). Although there are caveats in calculating variance explained by adding up single SNP main effects (for instance, it ignores possible interactions between markers), this approach reflects current practices in estimating variance for loci identified through large meta-analyses of genome-wide association study (GWAS) results. At the *BCL11A* locus, it is likely that the difference in phenotypic variance explained is due to the presence of HbF-increasing and HbF-decreasing alleles on the same haplotype background, where associated SNPs in LD masked each other's phenotypic effect (**Supplementary Table 5**). This antagonistic effect could represent an important source of the 'hidden' heritability highlighted by GWAS[7]. Imputation of ungenotyped markers did not reveal other SNPs with stronger association to HbF levels than rs10189857-rs7599488-rs4671393 (**Supplementary Table 6**).

The *HBS1L-MYB* intergenic interval carries DNA polymorphisms that influence HbF levels in healthy Europeans and in individuals of African ancestry with SCD[1,3,4]. We performed single-marker regression analysis and identified rs9402686, which was more strongly associated with HbF levels than the previous index HbF SNP at this locus ($P = 1.9 \times 10^{-13}$ for rs9402686 compared to $P = 3.5 \times 10^{-10}$ for rs9399137)[4] (**Table 1** and **Supplementary Table 3**). Stepwise conditional analysis uncovered two additional SNPs, ss244317976 and rs28384513, which were independently associated with HbF levels (**Table 1**). LD between rs9402686, ss244317976 and rs28384513 is weak ($r^2 < 0.03$). As for *BCL11A*, haplotypes defined by these three SNPs explained more variation in HbF levels than the cumulative sum of the phenotypic variance explained by the SNPs individually (7.3% compared to 6.8%), although the difference was not statistically significant (**Supplementary Table 7**).

In contrast to the *BCL11A* locus[5], we do not know the identity of the gene(s) that influence HbF levels in the *HBS1L-MYB* region.

[1]Montreal Heart Institute, Montréal, Québec, Canada. [2]Divisions of Genetics and Endocrinology and Program in Genomics, Children's Hospital Boston, Boston, Massachusetts, USA. [3]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [4]Division of Hematology and Oncology, Children's Hospital Boston, Boston, Massachusetts, USA. [5]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [6]Howard Hughes Medical Institute, Boston, Massachusetts, USA. [7]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [8]Département de Médecine, Université de Montréal, Montréal, Québec, Canada. Correspondence should be addressed to S.H.O. (stuart_orkin@dfci.harvard.edu), J.N.H. (joelh@broadinstitute.org) or G.L. (guillaume.lettre@mhi-humangenetics.org).

Received 13 July; accepted 13 October; published online 7 November 2010; doi:10.1038/ng.707

Table 1 Fetal hemoglobin association results in 1,032 individuals with SCA from the Cooperative Study of Sickle Cell Disease

Locus	SNP	Chr.	Position	EA/OA	EAF	β (s.e.)	P	Variance explained (%)	β (s.e.)	P	β (s.e.)	P
BCL11A							**Univariate analysis**		**Conditional on rs4671393**		**Conditional on rs4671393 and rs7599488**	
	rs4671393	2	60,574,455	A/G	0.27	0.604 (0.046)	3.7×10^{-37}	14.7				
	rs7599488	2	60,571,851	T/C	0.31	0.007 (0.046)	0.89	0.002	0.283 (0.046)	1.2×10^{-9}		
	rs10189857	2	60,566,739	G/A	0.31	–0.010 (0.046)	0.83	0.005	0.241 (0.046)	1.6×10^{-7}	–0.794 (0.223)	3.9×10^{-4}
HBS1L-MYB							**Univariate analysis**		**Conditional on rs9402686**		**Conditional on rs9402686 and ss244317976**	
	rs9402686	6	135,469,510	A/G	0.06	0.650 (0.087)	1.9×10^{-13}	5.1				
	ss244317976	6	135,470,367	G/A	0.02	0.567 (0.150)	1.6×10^{-4}	1.4	0.639 (0.146)	1.3×10^{-5}		
	rs28384513	6	135,417,902	G/T	0.21	–0.098 (0.054)	0.070	0.3	–0.162 (0.053)	0.0024	–0.174 (0.054)	0.0013
β-globin							**Univariate analysis**					
	rs10128556	11	5,220,259	T/C	0.10	0.421 (0.069)	1.3×10^{-9}	3.5				

Genomic positions are given according to NCBI build 36.1. The effect allele is on the forward strand. Effect size (β) and standard error (s.e.) are given in *z*-score units. Chr., chromosome; EA, effect allele; OA, other allele; EAF, effect allele frequency.

MYB is a transcriptional regulator of erythropoiesis, whereas *HBS1L* expression levels correlate with genotypes at HbF-associated SNPs[1]. In principle, one can establish causality by identifying rare and penetrant mutations in nearby candidate genes[8]. Resequencing 70 individuals with SCA identified one, six and four rare missense variants (minor allele frequency <1%) in *BCL11A*, *HBS1L* and *MYB*, respectively, that were absent from the 120 HapMap CEU and YRI samples. We genotyped these 11 rare variants in 1,032 individuals with SCA to assess their burden at the gene level by comparing normalized HbF levels in carriers and non-carriers (**Supplementary Methods**). To minimize ascertainment bias, we removed resequenced SCA cases from this analysis. This excluded singletons and left five and three variants to analyze in *HBS1L* and *MYB*, respectively. Results for *HBS1L* were not significant (corrected $P = 1$). However, we observed a significant difference for *MYB* (corrected $P = 0.005$), with the 25 carriers having on average 1.4% more HbF than the 937 non-carriers (**Table 2**). These data suggest that *MYB* is causally involved in controlling HbF production.

Recently, it has been suggested that some of the genetic associations identified by GWAS are due to collections of rare variants captured by common variants[9]. We tested whether the HbF association signals with common SNPs in the *HBS1L-MYB* intergenic region are due to the rare variants identified in *MYB*. LD between the three common SNPs and the three rare missense variants, as measured by D', is high ($r^2 < 0.01$, $D' > 0.4$; **Supplementary Table 8**). When we considered the three *MYB* missense variants as covariates, the association results between HbF levels and the three common *HBS1L-MYB* SNPs were not affected (**Supplementary Table 9**), indicating that 'synthetic associations' with rare markers in *MYB* cannot explain the HbF association signal in the *HBS1L-MYB* intergenic region. These results provide a clear example where both common and rare DNA sequence variants at the same locus are independently associated with the same phenotype.

The sickle cell mutation in the β-globin locus is associated with five 'classic' haplotypes (Benin, Bantu, Cameroon, Senegal and Arab-Indian) that are characterized by different degrees of clinical severity and HbF levels[10]. An *Xmn*I polymorphism (rs7482144) in the proximal promoter of *HBG2* marks the Senegal and Arab-Indian haplotypes and is associated with HbF levels in African Americans with SCD[4,11]. It remains unclear whether rs7482144-*Xmn*I is a causal variant at the β-globin locus. We replicated the association between rs7482144-*Xmn*I and HbF levels ($P = 3.7 \times 10^{-7}$) (**Supplementary Table 3**). However, rs10128556, located downstream of *HBG1*, was more strongly associated with HbF levels than rs7482144-*Xmn*I by two orders of magnitude ($P = 1.3 \times 10^{-9}$) (**Table 2**). When we conditioned on rs10128556, the HbF association result for rs7482144-*Xmn*I was not significant ($P = 0.78$ and $P = 0.047$ for rs10128556 when conditioned on rs7482144-*Xmn*I) (**Supplementary Fig. 2**). This indicates that rs7482144-*Xmn*I is not a causal variant for HbF levels in African Americans with SCA. Similarly, the recently described association between rs5006884 in the olfactory receptor gene cluster upstream of the β-globin genes and HbF levels was not significant after conditioning on rs10128556 ($P = 0.055$ and $P = 1.2 \times 10^{-6}$ for rs10128556 when conditioned on rs5006884) (**Supplementary Fig. 2**)[12]. Finally, when

Table 2 Role of rare missense DNA sequence variants in *HBS1L* and *MYB* on fetal hemoglobin levels

Gene	DNA sequence variant	MAF	Annotation	PolyPhen-2 prediction	Mean % HbF (carriers)	*n* (carriers)	Mean % HbF (non-carriers)	*n* (non-carriers)	*P*
HBS1L	ss212962438	0.0088	Arg44Trp	Probably damaging	5.83	17	6.08	948	–
	ss212962440	0.0021	Glu55Lys	Probably damaging	7.56	4	6.08	960	–
	ss212962441	0.0073	Ser65Cys	Benign	5.60	14	6.09	951	–
	ss212962478	0.0021	Asp13Glu	Probably damaging	7.96	4	6.06	955	–
	ss212962504	0.0010	Ser672Tyr	Benign	10.00	2	6.07	962	–
	All five HBS1L missense variants	–	–	–	6.09	40	6.07	917	1
MYB	rs73555746	0.0083	Glu626Ala	Probably damaging	7.87	15	6.05	949	–
	ss212962653	0.0005	Ser661Leu	Probably damaging	6.60	1	6.08	964	–
	ss212962648	0.0062	Gly628Glu	Benign	6.64	10	6.09	953	–
	All three MYB missense variants	–	–	–	7.47	25	6.06	937	0.005

Rare *HBS1L* and *MYB* missense variants with minor allele frequency (MAF) <1% were genotyped in 1,032 African Americans with SCA from the Cooperative Study of Sickle Cell Disease (CSSCD). We excluded 70 SCA cases used in the resequencing phase of this project from this analysis. The gene burden was assessed using Wilcoxon's rank test by comparing normalized HbF levels between carriers and non-carriers. *P* values are corrected for three genes tested (**Supplementary Methods**).

we conducted a haplotype analysis with the 43 SNPs genotyped at the β-globin locus and used rs10128556 as a covariate, the result was not significant ($P = 0.40$), indicating that rs10128556 (or a marker in LD with it) is the principal HbF-influencing variant at the β-globin locus in African Americans with SCA (**Supplementary Table 10**).

Studies of the genetic regulation of HbF have provided new biological insights: BCL11A maintains γ-globin silencing and is required for developmental switching within the β-globin cluster[5,6]. HbF-associated variants have also shown potential predictive value: these variants are associated with transfusion-independent β-thalassemia[3,13,14] and reduced pain crisis rate in SCD[4]. In this study, we showed that fine mapping of known associated loci through resequencing and dense genotyping can reveal additional independent association signals that could account for a significant fraction of the 'hidden' heritability[7]. For HbF levels, we increased the HbF phenotypic variation explained by the same three loci from 23.5% to 30.1%. Assuming a heritability of 60.9%, this translates to an increase from 38.6% to 49.5% of the heritable variation[15]. Thus, characterization of loci identified by GWAS will likely identify previously untested variants and explain part of the 'hidden' heritability for complex traits.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank all the individuals who participated in this study, and T. Nguyen and M. Beaudoin for DNA genotyping support. We thank S. Raychaudhuri for critical reading of the manuscript, G. Boucher for statistical advice and the CARe Sickle Cell Disease working group for providing the Cooperative Study of Sickle Cell Disease (CSSCD) principal components. This work was funded by the Fondation de l'Institut de Cardiologie de Montréal (to G.L.) and was supported by an Innovations in Clinical Research Award grant from the Doris Duke Charitable Foundation (to G.L. and J.N.H.). Resequencing services were provided by the University of Washington, Department of Genome Sciences, under US Federal Government contract number N01-HV-48194 from the National Heart, Lung, and Blood Institute.

AUTHOR CONTRIBUTIONS

V.G.S., S.H.O., J.N.H. and G.L. conceived and designed the experiment. G.G., C.D.P. and G.L. performed the experiments. G.G., C.D.P. and G.L. analyzed the data. G.G., C.D.P., V.G.S., S.H.O., J.N.H. and G.L. contributed reagents, materials and/or analysis tools. G.G. and G.L. wrote the paper with contributions from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Thein, S.L. *et al. Proc. Natl. Acad. Sci. USA* **104**, 11346–11351 (2007).
2. Menzel, S. *et al. Nat. Genet.* **39**, 1197–1199 (2007).
3. Uda, M. *et al. Proc. Natl. Acad. Sci. USA* **105**, 1620–1625 (2008).
4. Lettre, G. *et al. Proc. Natl. Acad. Sci. USA* **105**, 11869–11874 (2008).
5. Sankaran, V.G. *et al. Science* **322**, 1839–1842 (2008).
6. Sankaran, V.G. *et al. Nature* **460**, 1093–1097 (2009).
7. Manolio, T.A. *et al. Nature* **461**, 747–753 (2009).
8. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. *Science* **324**, 387–389 (2009).
9. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. *PLoS Biol.* **8**, e1000294 (2010).
10. Embury, S.H. *et al. Sickle Cell Disease: Basic Principles and Clinical Practice* (Lippincott Williams & Wilkins, Philadelphia, Pennsylvania, USA, 1994).
11. Labie, D. *et al. Proc. Natl. Acad. Sci. USA* **82**, 2111–2114 (1985).
12. Solovieff, N. *et al. Blood* **115**, 1815–1822 (2010).
13. Galanello, R. *et al. Blood* **114**, 3935–3937 (2009).
14. Nuinoon, M. *et al. Hum. Genet.* **127**, 303–314 (2009).
15. Pilia, G. *et al. PLoS Genet.* **2**, e132 (2006).