nature genetics

A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15

Stephan Menzel[1], Chad Garner[2], Ivo Gut[3], Fumihiko Matsuda[3], Masao Yamaguchi[3], Simon Heath[3], Mario Foglio[3], Diana Zelenika[3], Anne Boland[3], Helen Rooks[1], Steve Best[1], Tim D Spector[4], Martin Farrall[5], Mark Lathrop[3] & Swee Lay Thein[1,6]

F cells measure the presence of fetal hemoglobin, a heritable quantitative trait in adults that accounts for substantial phenotypic diversity of sickle cell disease and β thalassemia. We applied a genome-wide association mapping strategy to individuals with contrasting extreme trait values and mapped a new F cell quantitative trait locus to *BCL11A*, which encodes a zinc-finger protein, on chromosome 2p15. The 2p15 *BCL11A* quantitative trait locus accounts for 15.1% of the trait variance.

Genome-wide association methodology has recently identified susceptibility loci for several diseases, but it has a relatively high per-sample cost and requires large samples to detect modest risk effects. Strategies to increase power include selecting subjects with increased genetic load through early onset or identifying familial clustering of disease. Here, we apply a powerful alternative approach that uses a comparatively small number of study subjects taken from the extremes of a quantitative distribution.

In healthy adults, fetal hemoglobin (HbF; also known as $\alpha_2\gamma_2$) is present at residual levels ($<$0.6% of total hemoglobin) with over twenty-fold variation. Ten to fifteen percent of adults in the upper tail of the distribution have HbF levels between 0.8% and 5.0%, a condition referred to as heterocellular hereditary persistence of fetal hemoglobin (hHPFH)[1]. Although these HbF levels are modest in otherwise healthy individuals, interaction of hHPFH with β thalassemia or sickle cell disease (SCD) can increase HbF output in these individuals to levels that are clinically beneficial[2]. The ameliorating effect of HbF on SCD and β thalassemia has prompted numerous genetic and pharmacological approaches to reactivation of HbF synthesis[3], but the molecular mechanisms are not fully understood. Current pharmacological agents, such as hydroxycarbamide and butyrate analogs, show that it is possible to augment HbF production therapeutically, but these agents are limited by toxic effects and variable patient response.

HbF in the normal range (including hHPFH) is most sensitively measured by the proportion of F cells (that is, the proportion of erythrocytes containing measurable amounts of HbF[1]). The majority of the quantitative variation is highly heritable ($h^2 = 0.89$)[4], but the genetic etiology is complex, with several contributing quantitative trait loci (QTLs). To date, major QTLs have been identified with strong and reproducible statistical support at *Xmn*I-$^{G}\gamma$ in the β globin locus on chromosome 11p15 (ref. 5) and in the *HBS1L-MYB* intergenic region on chromosome 6q23 (ref. 6).

To map additional QTLs, we selected a panel of 179 unrelated individuals from the extreme upper and lower tails (above the 95th or below the 5th percentile points (that is, $>P_{95}$ or $<P_5$)) of the F cell distribution, drawn from a database of 5,184 phenotyped individuals from the St. Thomas Adult Twin Registry (http://www.twinsuk.ac.uk[7]), and genotyped them using the Illumina Sentrix HumanHap300 BeadChip (**Supplementary Methods** online). The study was approved by the local ethics committee of St. Thomas' and King's College Hospitals, London (LREC number 00-245), and all participants gave informed written consent. For the 308,015 markers retained after quality control, we assessed association using a Fisher exact χ^2 statistic for the allele counts in the high or low trait categories along with a linear regression analysis of the continuous trait against genotype (additive effects), with age and sex included as covariates. The two analyses gave similar results, and *P* values from the allele count test are presented in the text. Tests of non-additivity in the linear regression led to identical conclusions. Although extreme discordant sampling designs violate the usual normality assumption of linear regression, it does not inflate the type 1 error rate[8], which we confirmed by simulations and inspection of the Q-Q plot (**Supplementary Fig. 1** online). The genomic control parameter was 1.01, indicating that there was minimal admixture or cryptic relatedness in this sample[9]. Principal components analysis[10] confirmed this.

We identified major QTLs on chromosomes 2p15 ($P = 4.0 \times 10^{-16}$), 6q23 ($P = 8.8 \times 10^{-25}$) and 11p15 ($P = 1.7 \times 10^{-26}$) (**Fig. 1a**). The 6q23 QTL was first localized through linkage analysis in a large Asian-Indian family with beta thalassemia[11], then validated and fine-mapped in northern Europeans[6]. The association signal on 11p15 maps to the beta globin cluster, where the functional variant is thought to be the *Xmn*I-$^{G}\gamma$ variant at position –158 upstream of the $^{G}\gamma$ globin gene[5].

Markers within a 126-kb segment on chromosome 2p15 (nucleotides 60456396 to 60582798) identified a third, previously unreported QTL close to the oncogene *BCL11A*[12]. We genotyped an additional

[1]King's College London School of Medicine, Division of Gene and Cell Based Therapy, King's Denmark Hill Campus, London SE5 9PJ, UK. [2]University of California at Irvine, Epidemiology Division, Department of Medicine, Irvine, California 92697-7550, USA. [3]Centre National de Génotypage, Institut Génomique, Commissariat à l'Energie Atomique, 91006 Evry, France. [4]King's College London School of Medicine, Division of Genetics and Molecular Medicine, St. Thomas' Hospital, London SE1 7EH, UK. [5]The Wellcome Trust Centre for Human Genetics, Department of Cardiovascular Medicine, University of Oxford, Headington, Oxford OX3 7BN, UK. [6]King's College Hospital, Department of Haematological Medicine, Denmark Hill, London SE5 9RS, UK. Correspondence should be addressed to S.L.T. (sl.thein@kcl.ac.uk).

Received 15 March; accepted 2 July; published online 2 September 2007; doi:10.1038/ng2108

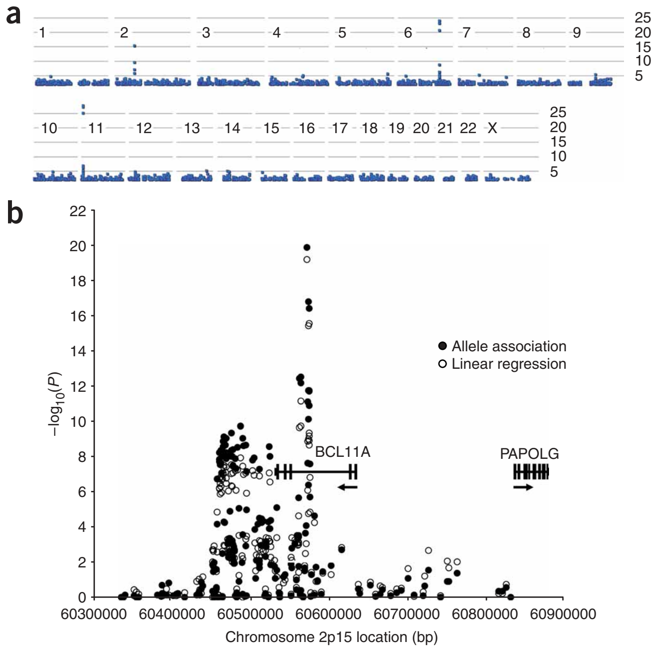

Figure 1 Association statistics ($-\log_{10}(P)$) for individuals included in the genome-wide screening panel. (**a**) Association statistics for 3,225 markers genome-wide with $P < 10^{-2}$. (**b**) Association statistics for 211 markers across the 2p15 region of association.

142 SNPs, 103 of which came from HapMap[13] and 39 of which were identified from dbSNP or by resequencing (**Supplementary Table 1** online). Analysis of this dense marker set uncovered two clusters of markers showing highly significant association at $P < 10^{-10}$ (**Fig. 1b**). The strongest associations (for example, $P < 10^{-19}$ at rs1427407) were in a region spanning 14 kb at nucleotides 60561398 to 60575745 in the second intron of *BCL11A*. The second association cluster spanned 67 kb at nucleotides 60457454 to 60523981 in the 3′ region of the gene, located approximately 8–74 kb downstream of exon 5. Markers that were significantly associated with the trait generally showed high LD within each cluster and lower LD between clusters (**Supplementary Fig. 2** online).

To corroborate our findings, we investigated two additional sample panels (the 'replication panel' and the 'twin panel', as defined below) with markers selected to represent the three QTLs (**Table 1**). For chromosome 2p15, we examined four markers from the first association cluster and two markers from the second association cluster. For 6q23, we chose markers from three linkage disequilibrium groups that contribute independently[6]. The *Xmn*I-$^{G}\gamma$ marker was genotyped on 11p15.

First, we replicated the associations in an independent group of 90 individuals with contrasting trait values (replication panel, $n = 90$, $<P_5$ or $>P_{95}$) (**Table 1**). Then, we measured the contribution of the marker to the overall trait variance in an unselected group of 720 twins ('unselected twins'; 310 dizygotic twin pairs and 100 monozygotic twin pairs) (**Table 1**). As related individuals were included, we applied a mixed linear model to test association and estimate residual heritability in the twin panel. The individual markers were all significantly associated with the trait (**Table 1**). A within-family test of association[14], which has less power but controls for possible population stratification, was significant for markers at the chromosome 2 and chromosome 6 QTLs. The trait variance attributed to each locus in the mixed linear model was 15.1% (95% confidence interval (c.i.) 12.6%–17.6%) for 2p15, 19.4% (16.6%–22.2%) for 6q22 and 10.2% (8.2%–12.2%) for 11p15. Tests of interactions between QTLs were nonsignificant, suggesting that they contribute additively. Together, they explain over 44% of the total trait variance in the twin panel (that is, half of the overall heritability of 89%).

Table 1 Results for representative markers for the three principal F cell QTLs

							Contributions to F cell variation (%)			
							Unselected twins N = 720			
		Allele frequency		Association test (*P* value)						
Polymorphism	Location (bp)	Unselected twins	Low / high F cell GWA and replication panels	GWA panel N = 179	Replication panel N = 90	Both N = 269	Variance (SNP)	ANOVA *P* value	QTDT[c] *P* value	Variance (locus)
Chr 2p15										
rs243027	60,460,511	0.43	0.44 / 0.72	4.6E-08	6.9E-04	2.2E-10	3.8	1.2E-04	n.s.	15.1
rs243081	60,467,280	0.48	0.41 / 0.71	3.8E-09	8.7E-05	2.5E-12	4.4	1.9E-04	n.s.	
rs6732518[a]	60,562,101	0.24	0.19 / 0.59	6.1E-13	2.0E-10	2.1E-21	11.1	1.1E-21	1.0E-04	
rs1427407[b]	60,571,547	0.14	0.03 / 0.42	2.5E-20	1.5E-11	6.1E-31	13.1	2.5E-22	1.7E-03	
rs766432	60,573,474	0.12	0.03 / 0.41	1.8E-17	5.8E-12	1.8E-28	13.5	1.7E-23	3.0E-04	
rs4671393[a,b]	60,574,455	0.12	0.03 / 0.41	5.5E-17	5.3E-11	2.6E-27	14.3	2.0E-22	9.0E-04	
Chr 6q23										
rs6904897[a]	135,424,673	0.37	0.35 / 0.56	8.2E-06	1.5E-02	1.2E-06	2.7	3.4E-05	8.0E-02	19.4
rs9399137[a]	135,460,711	0.23	0.18 / 0.57	2.8E-27	2.1E-11	2.5E-36	15.8	1.9E-30	6.0E-05	
rs1320963[a]	135,484,905	0.23	0.38 / 0.10	5.4E-10	1.2E-06	4.1E-15	6.7	9.0E-12	1.2E-02	
Chr 11p15.4										
*Xmn*I-$^{G}\gamma$[a]	5,232,745	0.33	0.10 / 0.63	2.0E-30	4.0E-11	2.4E-38	10.2	1.2E-17	n.s.	10.2

The within-family association test has been included for completeness. It is used principally in presence of population stratification (not found in our sample). It has less power than the ANOVA test and does not incorporate information on monozygotic twin pairs. n.s., not significant.
[a]Markers used to estimate the locus's contribution to the variance. [b]Markers not part of the genome-wide SNP set. [c]The within-family association test calculated with the QTDT program.

Haplotype analysis in the twin panel showed incomplete linkage disequilibrium, particularly between markers in the two association clusters (**Supplementary Tables 2** and **3** online)**.** A forward stepwise regression identified two markers (rs4671393 and rs6732518) from the first association cluster showing independent statistical effects on the trait. In particular, the markers from the second cluster did not show significant association after taking into account rs4671393 and rs6732518 (**Supplementary Table 4** online). This is consistent with the presence of more than one functional SNP or with the presence of untyped functional SNPs in incomplete LD with the typed markers from the first association cluster.

Accumulating experimental data are uncovering the genetic architecture of human quantitative variation. Resequencing studies of candidate genes in extreme groups have found diverse sets of rare, nonsynonymous alleles that collectively explain a modest proportion of the trait variance for some QTLs, whereas other QTLs are associated with common alleles—for example, circulating angiotensin 1 converting enzyme (ACE) activity. Applying GWA to individuals with contrasting extreme quantitative trait values is a powerful strategy for mapping common QTLs, as illustrated by our identification of three principal QTLs that contribute to F cells (and thus HbF).

One of the QTLs that we have identified is a new locus that maps to the gene encoding the C2H2-type zinc-finger protein BCL11A on chromosome 2p15, previously implicated in myeloid leukemia and lymphoma pathogenesis[12]. We examined multiple tissue cDNA panels by RT-PCR and found *BCL11A* to be expressed in a variety of tissues, including erythroid cells (**Supplementary Fig. 3** online). Mouse studies have shown that *Bcl11a* is essential for early lineage commitment in the development of T and B cells[12]. *BCL11A* has also been implicated in histone deacetylation and transcriptional repression in mammalian cells[15]. We speculate that dysregulated *BCL11A* expression may influence F cell production by affecting the kinetics of erythropoiesis[3].

It is likely that we have identified the principal QTLs that have frequent alleles affecting F cell production in the general European population, within the limits of the genome coverage of our markers. It is possible that additional loci could be uncovered with a denser map, but most of the remaining heritability is probably due to multiple small QTLs. The loci uncovered here have a major influence on the quantitative variation of the trait in healthy adults and possibly on the 'erythropoietic stress' responses underlying variability in β thalassemia and sickle cell disease severity and on the capacity of affected individuals to respond to pharmacologic inducers of HbF.

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

ACKNOWLEDGMENTS

We thank C. Steward for help in preparation of the manuscript. This work was supported by a grant from the UK Medical Research Council (MRC; G0000111 and ID 51640) to S.L.T. and by the French Ministry of Higher Education and Research (M.L.). Twins UK is supported by the Wellcome Trust and Framework V EU (European Union) grant 'Genome EU Twin'.

AUTHOR CONTRIBUTIONS

S.M. performed research, analyzed data and wrote the paper; C.G. analyzed data; M.Y. and M. Foglio performed bioinformatics analyses; I.G. and S.H. performed genome-wide association genotyping; D.Z., A.B., H.R., and S.B. performed research; T.D.S. contributed material; M. Farrall performed statistical genetic analysis and wrote the paper; M.L. codirected the research, analyzed data and wrote the paper; S.L.T. codirected the research and wrote the paper.

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. & Craig, J.E. *Hemoglobin* **22**, 401–414 (1998).
2. Steinberg, M.H., Forget, B.G., Higgs, D.R. & Nagel, R.L. (eds.). *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management* (Cambridge Univ. Press, Cambridge, 2001).
3. Bank, A. *Blood* **107**, 435–443 (2006).
4. Garner, C. *et al. Blood* **95**, 342–346 (2000).
5. Garner, C. *et al. GeneScreen* **1**, 9–14 (2000).
6. Thein, S.L. *et al. Proc. Natl. Acad. Sci. USA* **104**, 11346–11351 (2007).
7. Spector, T.D. & MacGregor, A.J. *Twin Res.* **5**, 440–443 (2002).
8. Tenesa, A., Visscher, P.M., Carothers, A.D. & Knott, S.A. *Behav. Genet.* **35**, 219–228 (2005).
9. Devlin, B. & Roeder, K. *Biometrics* **55**, 997–1004 (1999).
10. Patterson, N., Price, A.L. & Reich, D. *PLoS Genet.* **2**, e190 (2006) (doi:10.1371/journal.pgen.0020190).
11. Craig, J.E. *et al. Nat. Genet.* **12**, 58–64 (1996).
12. Liu, P. *et al. Nat. Immunol.* **4**, 525–532 (2003).
13. International HapMap Consortium. *Nature* **437**, 1299–1320 (2005).
14. Abecasis, G.R., Cardon, L.R. & Cookson, W.O. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
15. Senawong, T., Peterson, V.J. & Leid, M. *Arch. Biochem. Biophys.* **434**, 316–325 (2005).