A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression

*John J. Farrell,1 *Richard M. Sherva,1 Zhi-yi Chen,2 Hong-yuan Luo,2 Benjamin F. Chu,2 Shau Yin Ha,3 Chi Kong Li,4 Anselm C. W. Lee,5 Rever C. H. Li,5 Chi Keung Li,6 Hui Leung Yuen,7 Jason C. C. So,8 Edmond S. K. Ma,8 Li Chong Chan,8 Vivian Chan,9 Paola Sebastiani,10 Lindsay A. Farrer,11-13 Clinton T. Baldwin,1,14 Martin H. Steinberg,2 and David H. K. Chui2

Introduction

More than 3% of Chinese in Hong Kong are heterozygous carriers of β-thalassemia.1 Homozygotes or compound heterozygotes for β-thalassemia are usually severely ill and require monthly transfusions. Increased production of fetal hemoglobin (HbF; α2γ2) can modulate the disease severity by compensating for the shortfall of β globin caused by the β-thalassemia mutations.

HbF level in adults varies and is regulated as a multigenic trait.2 Three major HbF quantitative trait loci (QTL) have been identified: the C/T single nucleotide polymorphism (SNP, rs7482144) at promoter nucleotide (nt) 158 bp 5’ upstream of HBG2 on chromosome 11p15,3 the HBS1L-MYB intergenic polymorphism (HMIP) on chromosome 6q23,4 and the BCL11A polymorphism on chromosome 2p16.5 They can modulate HbF and disease severity in β-thalassemia,6-9 and sickle cell anemia.10 The relative contributions of these 3 QTLs to HbF regulation appear to differ among populations.5,11 The functional motif for each of these 3 QTLs responsible for their effects on HbF is not known.

In a genome-wide SNP association study (GWAS) among Chinese adult β-thalassemia heterozygotes in Hong Kong, both HMIP and BCL11A polymorphisms are highly associated with HbF level. To identify the functional variant in HMIP, we devised a novel strategy using the 1000 Genomes Project data, SNP imputation, variations between populations, and phylogenetic conservation to identify a hitherto unreported association between HbF expression and a 3-bp deletion polymorphism within the HMIP. Chromatin immunoprecipitation (ChIP) assays revealed binding of erythropoiesis-related transcription factors near the polymorphism. The DNA fragment surrounding the 3-bp deletion polymorphism has enhancer-like activity that is further augmented by the introduction of the 3-bp deletion. This 3-bp deletion polymorphism is probably the most significant functional motif accounting for HMIP modulation of HbF in all 3 populations. (Blood. 2011;117(18):4935-4945)

Methods

Study subjects

Chinese β-thalassemia adult carriers were parents of β-thalassemia major or intermedia patients at the Queen Mary, Prince of Wales, Tuen Mun, and Africans. Chromatin immunoprecipitation assays confirmed erythropoiesis-related transcription factors binding to this region in K562 cells. Based on transient expression of a luciferase reporter plasmid, the DNA fragment encompassing the 3-bp deletion polymorphism has enhancer-like activity that is further augmented by the introduction of the 3-bp deletion. This 3-bp deletion polymorphism is probably the most significant functional motif accounting for HMIP modulation of HbF in all 3 populations.
Queen Elizabeth, and Princess Margaret Hospitals in Hong Kong. All subjects signed informed consent in accordance with the Declaration of Helsinki. Identifying information was removed from data files prepared for analyses. An additional 300 archived genomic DNA samples from unrelated adult β-thalassemia heterozygotes at the Queen Mary Hospital were also studied. This research was approved by the Institutional Review Boards of the Boston University School of Medicine and each of the 5 hospitals in Hong Kong.12

African American subjects with sickle cell anemia were from the Cooperative Study of Sickle Cell Disease as reported by Solovieff et al.11

Hematology and hemoglobin analyses
Peripheral blood samples anticoagulated with ethylenediaminetetraacetic acid were delivered within 1 day after phlebotomy to the Division of Hematology, Department of Pathology, Queen Mary Hospital for clinical laboratory testing.12 F-cell measurement by flow cytometry based on detection by anti-Hb F antibody was previously described.12

Globin genotyping
Genomic DNA was extracted from peripheral blood leukocytes. β-Thalassemia mutations and XmnI polymorphism in HBG2 promoter were determined as described.12

SNP genotyping and quality control procedures
Genome-wide genotyping was performed on 659 subjects using the Illumina Human 610-Quad BeadChip array according to the Illumina Infinium II Assay protocol. A total of 500 ng of genomic DNA was used per sample. Genotypes were called by the Illumina BeadStudio Genotyping Module using Illumina supplied predetermined clusters for each SNP. Quality checks on the 582 539 genotyped SNPs completed on 659 subjects were performed to identify SNP call rates < 95%, SNPs with a minor allele frequency (MAF) < 1%, and SNPs not in Hardy-Weinberg equilibrium at a cut-off of P < .001. These procedures identified 17 282 SNPs with call rates below the threshold, 85 091 SNPs with a low MAF, and 5797 SNPs not in Hardy-Weinberg equilibrium. These 108 170 SNPs were excluded from the analysis. In addition, 32 persons with missing data for > 8% of the SNPs and 8 persons with extreme HbF values were excluded. The latter included subjects who were heterozygous for (δβ)0-thalassemia, (δγβ)0-thalassemia, (δγβ)δ-thalassemia, or HPFH deletions. For the remaining 619 subjects, the total genotyping rate was 99.2%.

Statistical analysis
HbF level was log-transformed to normalize the distribution. Tests for association between SNPs and HbF level were conducted using linear regression models implemented in PLINK, Version 1.05 (www.pmgu.mgh.harvard.edu/purcell/plink). Regression models were adjusted for the presence of the promoter nt A28 G or nt A29 A > G β -thalassemia mutations. Using a Bonferroni correction to adjust for multiple testing, a P value cutoff of 1.05 × 10−7 was set to establish genome-wide significance.

SNP imputation and analyses
SNP imputation within the HMIP region was carried out using the software program MACH, Version 1.0.13,14 together with the 1000 Genomes Project phased data released in August 2009 (www.1000genomes.org). For SNP imputation in Chinese adult β-thalassemia heterozygotes, a reference panel of 120 haplotypes from unrelated HapMap Chinese/Japanese (CHB/JPT) persons was used. For imputation in African Americans with sickle cell anemia, a merged reference panel of 242 haplotypes from unrelated HapMap White and Yoruban (YRI) persons was used. Only those imputed SNPs with an imputation quality of r2 more than 0.30 and MAF more than 0.01 were included for further analyses. Association analyses were performed using linear additive models of the normalized percentage HbF and the imputed allele dosages. Regression analyses were completed with the lm function available in the R software package Version 2.9.2 (cran.r-project.org). Models for the Chinese adult β-thalassemia heterozygotes were adjusted for the presence of promoter nt A28 and A29 β -thalassemia mutations.

Figure 1. Manhattan plot showing the GWAS results of 619 unrelated adult β-thalassemia heterozygotes in Hong Kong, using the additive model. The association was with normalized HbF expressed as percentage determined by HPLC.
Amplification refractory mutation system test for the 3-bp deletion

A bidirectional polymerase chain reaction (PCR) test using 2 pairs of allele-specific oligonucleotide primers was designed to detect the TAC 3-bp deletion on chromosome 6q23. The primer pairs were TAC-1 5'-H11032-TCACTCTGGACAGCAGATGTTACTAT-3'; and TAC-2, 5'-H11032-CTCAGTGATGGTATTTCTGGAGAC-3' to detect the sequence with intact TAC with a PCR product of 207 bp; TAC-3, 5'-H11032-AGCCCGTCCAGACACTCAT-TGTT-3'; and TAC-4, 5'-H11032-GCCCTGATAACATTTTGTGGTTTTCATTTAA-CAT-3' to detect the sequence with the 3-bp deletion with a PCR product of 276 bp.

PCR reaction was carried out in 10 µL containing 50 ng of target DNA; 5 µL of multiplex PCR master mix (QIAGEN), 1 µL of Q solution, 0.5 µL of each primer (125nM TAC-1, 500nM TAC-2, 62.5nM TAC-3, 500nM TAC-4), using GeneAmp PCR System 9700 (Applied Biosystems) at 98°C for 15 minutes, followed by 30 cycles of 98°C for 20 seconds, 58°C for 40 seconds, and 72°C for 40 seconds, and ended with the last cycle of 72°C for 10 minutes. A 5-µL aliquot of the PCR reaction was electrophoresed on 2% agarose gel.

Test for phylogenetic conservation

Nucleotide sequences were examined for conservation with 44 vertebrate species using the Phylop and PhastCon software analyses available through the University of California Santa Cruz (UCSC) Genome Browser tracks.15

ChIP assay

The SimpleChIP Enzymatic Chromatin IP Kit with magnetic beads (Cell Signaling Technology) was used according to the manufacturer’s protocol. Briefly, K562 cells were treated with formaldehyde (final concentration 1%) for 10 minutes at room temperature. Glycine (final concentration 125mM) was added to quench cross-linking. Cells were washed with ice-cold phosphate-buffered saline, pelleted by centrifugation at 4°C, and were lysed (buffer A) for 10 minutes on ice. Nuclei released were resuspended in ice-cold buffer B and digested with micrococcal nuclease (5 µL, 2000 gel units/µL) for 20 minutes at 37°C. The enzyme reaction was stopped by ethylenediaminetetraacetic acid, and the sheared chromatin was collected by centrifugation for 10 minutes at 4°C. A total of 5 µg of antibodies each against TAL1 (04-123, clone BTL73, Millipore), E47 (554077, clone G127-32, BD Bioscience), GATA-1 (SC-266, clone N1, Santa Cruz Biotechnology), GATA-2 (MAB2046, clone 527530, R&D systems), RUNX1 (ab23980, polyclonal, Abcam), and histone H3 (D2B12, Cell Signaling Technology) as well as normal IgG control were used for immunoprecipitation with 10 µg of sheared chromatin.

Figure 2. Effect of BCL11A and HMIP QTL upon HbF and F cells. (A) Box plots showing the complete distribution of HbF in 619 Chinese β-thalassemia heterozygotes, expressed as percentage in log scale on the y-axis. Each rectangle represents the data between the 25th and 75th quartiles, and the bar in each rectangle is the median value for the HbF in log scale. rs766432: AA, AC, and CC represent the SNP genotypes for rs766432 in BCL11A on chromosome 2p16. rs9399137: TT, CT, and CC represent the SNP genotypes for rs9399137 in HMIP on chromosome 6q23. (B) Effect of the 3 major HbF QTLs on HbF and F-cell levels in adult β-thalassemia heterozygotes who had HbF and F cells determined. Data are mean ± SD. Control indicates 80 subjects (age, 48.3 ± 9.0 years) who are homozygous for the major alleles for rs7482144 in HBG2 promoter, rs766432 in BCL11A, and rs9399137 in HMIP. rs766432: 8 subjects (age 46.9 ± 9.6 years) with identical HbF QTL genotypes as those in control, except that they are homozygous for the minor allele for rs766432 in BCL11A. rs9399137: 6 subjects (age 40.3 ± 13.8 years) with identical HbF QTL genotypes as those in control, except that they are homozygous for the minor allele for rs9399137 in HMIP.
After overnight incubation at 4°C, ChIP-grade protein G magnetic beads were added and incubated for 2 hours at 4°C. Protein G magnetic bead pellet was washed with low- and high-salt ChIP buffer. Cross-linking was reversed, and chromatin DNA was eluted with DNA elution buffer. Real-time PCR with Power SYBR Green PCR Master Mix (Applied Biosystems) using primer set, forward (5'-H11032-ATTCACTCTGGACAGCAGAT-GTA-3'
/H11032) from chromosome 6, 135 460 304 to 135 460 327 and reverse (5'-H11032-CCAGTAAGTGTCTTCTGAGGGAAC-3'/H11032) from chromosome 6, 135 460 383 to 135 460 360 was performed. The amount of immunoprecipitated DNA in each sample was determined as a fraction relative to input chromatin in percentages. The assay with histone H3 antibody served as a positive control using SimpleChIP Human RPL30 Exon 3 primer set.

For ChIP assays at a region approximately 1.5 kb centromeric to the 3-bp deletion polymorphism, the primer sets are forward (5'-ATTCACTCTGGACAGCAGAT-GTA-3') from chromosome 6, 135 460 304 to 135 460 327 and reverse (5'-CCAGTAAGTGTCTTCTGAGGGAAC-3') from chromosome 6, 135 460 383 to 135 460 360 was performed. The amount of immunoprecipitated DNA in each sample was determined as a fraction relative to input chromatin in percentages. The assay with histone H3 antibody served as a positive control using SimpleChIP Human RPL30 Exon 3 primer set.

For ChIP assays at a region approximately 2.5 kb telomeric to the 3-bp deletion polymorphism, the primer sets are forward (5'-CTGGAGTACGTGGTGAGTCTTG-3') from chromosome 6, 135 462 802 to 135 462 825 and reverse (5'-CACACCGTGAATCCGCTTGG-3') from chromosome 6, 135 462 903 to 135 462 880.

Cell culture, plasmid constructs, and site-directed mutagenesis

Culture of human erythroleukemia K562 cells and plasmid construct of luciferase reporter driven by 1.4 kb human HBG2 globin gene promoter (pGL3-1383/49 Luc) were previously described. The 61-bp wild-
type and 3-bp deletion fragments (see Figure 3A) were generated by PCR and cloned into the Kpn I and Spe I sites of the pGL3 WT-1383/49G/Luc (upstream of the 1.4 kb HBG2 globin gene promoter); the pGL3 WT-1383/49G/Luc and pGL3 Del-1383/49G/Luc plasmids. These constructs were sequenced to confirm their nucleotide fidelity. Mutations in the E-box (from CAGATG to TTTTAT), 5′GATA (from TATC to AAAA), RUNX1 binding (from ACCACA to TTTTTT), and 3′GATA (from TATC to AAAA) binding sites of 61 bp WT fragment were introduced into either plasmid using the QuikChange II site-directed mutagenesis kit (Stratagene) according to the manufacturer’s protocols. All mutations were verified by nucleotide sequence analysis.

**Transient transfection and luciferase assays**

K562 cells were seeded at 1.0 to 2.0 × 10^5 cells/well in 24-well tissue culture plates. The transfection and luciferase assay were as previously described, except that 0.25 μg of plasmid DNA was used in each transfection.

**Results**

**Genome-wide SNP association study**

A GWAS for HbF level was performed among 619 unrelated Chinese adult β-thalassemia heterozygotes, and the results are shown in the Manhattan plot (Figure 1). The highly significant SNPs in HMIP and BCL11A are tabulated in supplemental Table 1 (available on the Blood Web site; see the Supplemental Materials link at the top of the online article). The most significant SNP in HMIP was rs9399137 (P = 1.39 × 10^-23) and in BCL11A was rs766432 (P = 2.40 × 10^-15). SNP rs7482144 was not interrogated by the Illumina Human 610-Quad BeadChip array and was not analyzed in this GWAS. The linkage disequilibrium (LD) plots for the HMIP on chromosome 6q23 and BCL11A region on chromosome 2p16 generated using Haploview Version 3.2 are shown in supplemental Figure 1A and B.

**Effect of BCL11A and HMIP on HbF**

The distribution of log-transformed HbF values according to SNP genotypes for rs766432 in BCL11A and rs9399137 in HMIP are shown in box plots (Figure 2A). For both SNPs, homozygotes for the major allele have the lowest median HbF, whereas homozygotes for the minor allele have the highest median HbF. The variance in HbF level because of rs766432 was 6.4%; that because of rs9399137 was 13.5%. Tests to detect interaction between these 2 SNPs were not significant (P > .05), suggesting that their effects on HbF are independent.

Among subjects in whom both HbF and F cells (erythrocytes containing HbF) were determined, 80 were found to be homozygotes for the major alleles of SNPs in all 3 major QTLs (labeled as control in Figure 2B). Homozygotes for the major allele of rs766432 in BCL11A had twice the HbF and F cells compared with control. Homozygotes for the minor allele of rs9399137 had even higher HbF and F cells (Figure 2B).

**Mining data from the 1000 Genomes Project**

To identify the functional variant in HMIP that accounts for its association with HbF expression, we first interrogated the 1000 Genomes Project database (www.1000genomes.org) for all known polymorphisms within HMIP region spanning approximately 126 kb from chromosome 6: 135 417 715 to 135 544 146 bp. In the 59 CHB/
JPT, 57 CEU, and 56 YRI HapMap samples, which were sequenced for the 1000 Genomes Project, there were 478 SNPs in the CHB/JPT, 323 in the CEU, and 633 in the YRI samples. Of the SNPs found in CHB/JPT and CEU subjects, 235 were present in both populations. Twenty-seven of these are in strong LD (Pearson correlation $r^2 > 0.54$) with rs9399137 (supplemental Table 2).

We searched for insertion or deletion polymorphisms in the HMIP-2 block, approximately 24 kb from chromosome 6: 135 452 921 to 135 477 194.4 in the 1000 Genomes Project database and found a TAC 3-bp deletion between chromosome 6: 135 460 326 and 135 460 328 (Figure 3A) that is in very strong LD with rs9399137 in Chinese, Europeans, and African populations. SNP rs7775698 is located at chromosome 6: 135 460 328 bp, with CT being its major/minor alleles, respectively. In Chinese and European populations, the T allele as determined by the probe for rs7775698 primarily tags the 3-bp deletion (supplemental Table 3). In African populations, the T allele only very occasionally tags the 3-bp deletion. In contrast, the T allele in African populations primarily tags an ancestral sequence containing rs7775698 (T) allele without the 3-bp deletion (supplemental Table 3). This rs7775698 (T) allele without the 3-bp deletion found in African populations is also the reference allele found in several simian primates, such as chimpanzee, gorilla, orangutan, and Rhesus monkey. The frequency of this ancestral sequence in African populations is comparable with the frequency of the sequence with the 3-bp deletion found in Chinese and European populations.

The rs7775698(T)-rs9399137(C) haplotype that tags the 3-bp deletion is very common in both the CHB and CEU 1000 Genomes Project samples with frequency of 31.5% and 21.4%, respectively. In the YRI population, this 3-bp deletion haplotype is much less common with a frequency of 5.1%. To examine the difference between non-African and African populations, we calculated the $r^2$ between rs7775698 and rs9399137 in the 11 HapMap populations. All the non-African populations had a mean $r^2$ (0.94) much higher than that in African populations (0.17) (supplemental Table 4). The frequency of the rs7775698(T)-rs9399137(C) haplotype is 0.23 in non-African HapMap populations but only 0.05 in African HapMap populations.

The difference in the $r^2$ is the result of the presence of the ancestral sequence containing rs7775698 (T) without the 3-bp deletion that is present in 17% of the African populations but is rarely found in non-African populations ($< 0.01$). The rs9399137 (C) tags the sequence containing rs7775698 (T) with the 3-bp deletion in the Chinese and European populations examined. As a result of this LD pattern between rs9399137 (C) and the 3-bp deletion, the deletion is as highly significantly associated with HbF as the nearby (383 bp) rs9399137 (C) allele (see “Confirmation of the 3-bp deletion in study cohort”).

**Confirmation of the 3-bp deletion in study cohort**

To determine whether our Chinese study cohort had the 3-bp deletion in LD with rs9399137, we sequenced this region in 36 subjects who were homozygous for either the major allele T (n = 16) or minor allele C (n = 20) at rs9399137. All 16 T/T homozygotes are homozygous for having intact TAC in their nucleotide sequences (Figure 3B upper figure), and their HbF was 0.85% ± 0.70%. All 20 C/C homozygotes are homozygous for the TAC 3-bp deletion (Figure 3B lower figure), and their HbF was 3.24% ± 2.09%.

A PCR-based amplification refractory mutation system test capable of detecting homozygosity, heterozygosity, and wild-type for the 3-bp deletion was designed. This was applied to 335 unrelated Chinese β-thalassemia heterozygotes. There were 17 homozygotes for the 3-bp deletion, 115 heterozygotes, and 203 wild-type. The 3-bp deletion frequency is 0.22. Furthermore, the 3-bp deletion is in complete LD with rs9399137, the SNP found in GWAS to be most significantly associated with HbF. Both polymorphisms were equally associated with HbF with an identical $P$ value of $1.0 \times 10^{-20}$ among these 335 subjects.

**Evidence for the functionality of the TAC 3-bp deletion**

1. **Comparison of association results between populations.** Under the assumption that the functional variant should be the most strongly associated with HbF in all populations, we sought to leverage the cross-population differences in LD patterns and allele frequencies in the region to narrow the field of candidates. We imputed SNPs in our study cohort of 619 Chinese β-thalassemia heterozygotes and in 848 African Americans with sickle cell anemia11 using 1000 Genomes Project haplotype reference panels. Results of the association analyses of the imputed SNPs with HbF in Chinese subjects are shown in Figure 4. The most significant SNP is rs9399137 ($P = 7.34 \times 10^{-25}$). All other SNPs with significant HbF association are also in LD with rs9399137.

A total of 27 genotyped and imputed SNPs were found to be in strong LD with rs9399137 and highly associated with HbF in the Chinese β-thalassemia heterozygotes (supplemental Table 2). Among African Americans, 14 of these SNPs have MAFs greater than 0.17 yet are only weakly associated with HbF, with rs4895440 being the most significant ($P = 4.93 \times 10^{-4}$). One other SNP (chromosome 6q23 135 474 996 bp) is monomorphic in African Americans. Given the weak association results with these SNPs and the fact that these SNPs had significantly high minor allele frequencies to afford sufficient power to detect association with HbF if present, these 15 SNPs were excluded from consideration as probably functional motifs. Two more SNPs (rs9494145 and rs9483788) were also removed from consideration because they were substantially less strongly associated with HbF ($P = 1.73 \times 10^{-18}$ and $4.59 \times 10^{-16}$, respectively) than rs9399137 ($P = 7.34 \times 10^{-25}$) in the Chinese samples. The remaining 10 SNPs and the 3-bp deletion polymorphism merit further study in the search for the functional variant (supplemental Table 5).

2. **Transcription factor binding sites and phylogenetic conservation.** We next assessed the nucleotide sequences near each of the 11 polymorphisms for potential transcription factor binding sites and phylogenetic conservation. We searched for insertion or deletion polymorphisms in the HMIP-2 block, approximately 24 kb from chromosome 6: 135 452 921 to 135 477 194.4 in the 1000 Genomes Project database and found a TAC 3-bp deletion between chromosome 6: 135 460 326 and 135 460 328 (Figure 3A) that is in very strong LD with rs9399137 in Chinese, Europeans, and African populations. SNP rs7775698 is located at chromosome 6: 135 460 328 bp, with CT being its major/minor alleles, respectively. In Chinese and European populations, the T allele as determined by the probe for rs7775698 primarily tags the 3-bp deletion (supplemental Table 3). In African populations, the T allele only very occasionally tags the 3-bp deletion. In contrast, the T allele in African populations primarily tags an ancestral sequence containing rs7775698 (T) allele without the 3-bp deletion (supplemental Table 3). This rs7775698 (T) allele without the 3-bp deletion found in African populations is also the reference allele found in several simian primates, such as chimpanzee, gorilla, orangutan, and Rhesus monkey. The frequency of this ancestral sequence in African populations is comparable with the frequency of the sequence with the 3-bp deletion found in Chinese and European populations.

The rs7775698(T)-rs9399137(C) haplotype that tags the 3-bp deletion is very common in both the CHB and CEU 1000 Genomes Project samples with frequency of 31.5% and 21.4%, respectively. In the YRI population, this 3-bp deletion haplotype is much less common with a frequency of 5.1%. To examine the difference between non-African and African populations, we calculated the $r^2$ between rs7775698 and rs9399137 in the 11 HapMap populations. All the non-African populations had a mean $r^2$ (0.94) much higher than that in African populations (0.17) (supplemental Table 4). The frequency of the rs7775698(T)-rs9399137(C) haplotype is 0.23 in non-African HapMap populations but only 0.05 in African HapMap populations.

The difference in the $r^2$ is the result of the presence of the ancestral sequence containing rs7775698 (T) without the 3-bp deletion that is present in 17% of the African populations but is rarely found in non-African populations ($< 0.01$). The rs9399137 (C) tags the sequence containing rs7775698 (T) with the 3-bp deletion in the Chinese and European populations examined. As a result of this LD pattern between rs9399137 (C) and the 3-bp deletion, the deletion is as highly significantly associated with HbF as the nearby (383 bp) rs9399137 (C) allele (see “Confirmation of the 3-bp deletion in study cohort”).

**Figure 5. ChIP assay.** Relative real-time PCR signal intensities of the ChIP assay, showing significant binding of TAL1, E47, GATA2, and RUNX1 near the 3-bp deletion polymorphism. ■ represents transcription factor binding signals in the immediate vicinity of the 3-bp deletion polymorphism; ■ binding signals at approximately 1.5 kb centromeric to the 3-bp deletion polymorphism; and ■ binding signals at approximately 2.5 kb telomeric to the 3-bp deletion polymorphism. IgG indicates negative control using IgG for immunoprecipitation; H3, positive control using antibody against histone H3; and RT-PCR, using exon 3 primer set.
sites using the transcription factor binding site search program ConSite (www.consite.genereg.net). The transcription factor binding site scores are based on a profile model that uses position specific weight matrices to assign a score to the candidate sequences in conserved regions. The results are presented in supplemental Table 5. The region flanking the 3-bp deletion is of particular importance because it provides a potential platform for the binding of 4 transcriptional factors known to be essential for erythroid cell differentiation and hemopoiesis: TAL1/SCL, E47, GATA, and RUNX1/AML1 (Figure 3A).

The sequences flanking the 3-bp deletion polymorphism are highly conserved among 44 vertebrate species with log odds conserved element scores of 64 and 84 (supplemental Figure 2). Sequences near rs11154792 are conserved with a lesser log odds score of 17. Sequence conservation was not found in regions near the other variants.

3. Transcription factor binding. Based on nucleotide sequencing results, K562 cells are heterozygous for the 3-bp deletion polymorphism. ChIP assay using K562 cells revealed that GATA-2 binds to the immediate vicinity of the 3-bp deletion polymorphism, the signal being 5- to 15-fold higher compared with binding at sites 1.5 kb telomeric or 2.5 kb centromeric to the 3-bp deletion polymorphism, respectively (Figure 5; supplemental Figure 3). The signal for GATA-1 binding, if present at all, is much less than GATA-2. These results are consistent with a published report on ChIP-seq experiments on genome-wide GATA binding site occupancy. One GATA-2 ChIP-seq peak, but not GATA-1, was located between chromosome 6: 135 460 050 to 135 460 379, which encompasses the 3-bp deletion polymorphism under study at chromosome 6: 135 460 326 to 135 460 328. No GATA ChIP-seq peak is found near the rs9399137 located at chromosome 6: 135 460 711. (See also ENCODE data later in Results.)

Significant binding of TAL1, E47, and RUNX1 to the DNA region surrounding the 3-bp deletion polymorphism is also shown by ChIP assays in K562 cells (Figure 5). E47 binding is 2- to 7-fold higher compared with binding signal at sites 1.5 kb telomeric or 2.5 kb centromeric to the 3-bp deletion polymorphism. Less difference is observed with the TAL1 binding. No difference is observed with the RUNX1 binding.

The ENCODE datasets provide a rich resource of experimental results showing potential regulatory regions throughout the genome. Figure 6 shows that, within the immediate vicinity of the 3-bp deletion polymorphism in K562 cells, ChIP-seq experiments revealed GATA-2 occupancy, less so for GATA-1, and occupancy of BRG1, INI1 also known as BAF47, and RNA polymerase II. Present also are the digital DNase genomic footprinting and RNA transcripts. There is a strong signal for ESPERR regulatory potential in this region among several species. In contrast, these transcription regulatory markers are either absent or present at a low level in the immediate vicinity of rs9399137 (Figure 6). The ChIP-seq signals for histones H3K4Me1 and H3K4Me3 in the vicinity of the 3-bp deletion polymorphism and rs9399137 region are relatively low, and there appears to be no significant enrichment between these 2 sites.

4. Enhancer-like activity. In an initial attempt to demonstrate functionality of the DNA fragment surrounding the 3-bp deletion...
polymorphism, DNA fragment either without or with the 3-bp deletion (Figure 3A) was ligated to an expression vector consisting of 1.4 kb HBG2 proximal promoter and luciferase reporter gene. These were transiently transfected into K562 cells, and the luciferase activity was measured after 48-hour cell culture.

The 61-bp DNA fragment without the 3-bp deletion enhances the HBG2 promoter transcriptional activity by 3.4-fold, compared with the plasmid without the 61-bp DNA fragment (Figure 7A). The 58-bp DNA fragment with the 3-bp deletion enhances the HBG2 promoter transcriptional activity by 5.4-fold, compared with the plasmid without the 58-bp DNA fragment (Figure 7A).

The enhancer-like activity of the 61-bp DNA fragment is dependent on intact RUNX1 and 3′/H11032 GATA binding sites. Mutation of either of these 2 binding sites down-regulates the enhancer-like activity of the DNA fragment to approximately 50%. Mutation of the E-box or 5′/H11032 GATA binding site does not decrease the enhancer-like activity (Figure 7B). Mutation of both GATA sites results in lower enhancer-like activity, suggesting that the 3′/H11032 GATA site rather than the 5′/H11032 GATA site has an important role in modulating the enhancer-like activity (Figure 7B). The 3′/H11032 GATA site (TGATAA) matches well with the WGATAR sequence found
most commonly in the DNA segments occupied by GATA in erythroid cells. On the other hand, the 5' GATA site (TGATAT) is not as good a match to the consensus sequence.

The enhancer-like activity of the 58-bp DNA fragment with the 3-bp deletion is dependent on intact E-box, RUNX1, and 3' GATA binding sites. Mutation of any of the 3 binding sites significantly down-regulates the enhancer-like activity of the DNA fragment to approximately half of the intact and unmuted DNA fragment with the 3-bp deletion (Figure 7C).

Discussion

The HbF QTL on chromosome 6q23 was first identified in an Asian Indian family segregating with β-thalassemia and hereditary persistence of fetal hemoglobin.25 This locus is now localized within HMIP.4 HMIP also impacts on erythrocyte, platelet, monocyte numbers, and erythrocyte parameters, such as hematocrit, mean cell volume, mean cell hemoglobin, and mean cell hemoglobin concentration.26-28

HMIP and the BCL11A polymorphism are highly associated with HbF in Chinese adult β-thalassemia carriers.18,29 Although GWASs have robustly identified genes and loci associated with diseases or phenotypes, identification of responsible functional variants has proven difficult. The 1000 Genomes Project (www.1000genomes.org) provides a large number of publicly available genomic sequences from different populations that can be interrogated to identify common and rare variants within haplotype blocks of interest.30,31 Lists of these candidate variants can be further refined based on bioinformatics and variations among populations. Using this strategy, we identified a hitherto unreported association between HbF expression and a 3-bp deletion polymorphism in HMIP. A recent study resequenced 32.8 kb of the HbF polymorphism has enhancer-like activity, which is substantiated by ChIP-seq results that, in the same region, there is GATA-2 binding, occupancy of BRG1, IN11, RNA polymerase 2, and to a lesser extent histones H3K2Me1 and H3K2Me4, as well as footprinting (Figure 6). The presence of 50-bp RNA transcripts near the 3-bp deletion polymorphism is also suggestive of the presence of an enhancer.22,23

We hypothesize that the region surrounding the 3-bp deletion polymorphism has enhancer-like activity, which is substantiated by results of transfection experiments using the HBG2 1.4-kb proximal promoter linked to a luciferase reporter gene. Without the deletion, the DNA fragment enhances the HBG2 promoter activity by 3.4-fold (Figure 7A). With the 3-bp deletion, the promoter activity enhancement increased to 5.4-fold (Figure 7A). Mutating the various transcription factor binding sites led to down-regulation of the enhancer-like activity (Figure 7B-C), indicating that transcription factor binding is necessary for the enhancer-like activity. These are initial attempts to demonstrate functionality of the DNA fragment surrounding the 3-bp deletion polymorphism. Further investigations are needed to ascertain in vivo biologic function of this DNA fragment in primary human adult erythroid cell cultures with or without the 3-bp deletion, and to identify its target genes for the enhancer-like activity.

The 3-bp deletion polymorphism is located at 42.6 kb upstream of HBS1L and 83.8 kb upstream of MYB, near the erythroid specific Dnase hypersensitive site 2 within the HMIP-2 block.42 MYB protein can increase erythroid cell proliferation and inhibit cellular differentiation. Overexpression of MYB, but not HBS1L, in human K562 erythroid cells was reported to decrease HBG expression and supports the hypothesis that MYB modulates HbF production by altering erythropoiesis kinetics.43 Down-regulation of MYB expression by shRNAs in adult erythroid cell cultures was reported to result in increased HBG expression.44 On the other hand, HBS1L, but not MYB, expression was found to correlate with HMIP alleles associated with high HbF in primary human adult erythroid cell cultures.4 HBS1L has 4 guanosine triphosphate-binding motifs, but its biologic functions are not understood. It remains undetermined whether HMIP QTL modulates HbF production primarily
through MYB, HBSIL, both, or neither. Our transfection experiments carried out in K562 cells showed that the 58-bp DNA fragment with the 3′-bp deletion, representing the minor allele of chromosome 6q23 QTL, which is associated with elevated HbF, has increased enhancer-like activity. These observations suggest that the function of this HbF QTL is not necessarily mediated through direct transcriptional regulation of MYB. Near the QTL, there is PDE7B coding for a cyclic adenosine 5′-monophosphate-dependent phosphodiesterase which can effect HBG expression and erythroid cell differentiation through cyclic adenosine 5′-monophosphate pathway.46 Located further telomeric is MAP3K5.47 Histone hyperacetylation by activation of p38 MAP kinase pathway is linked to HBG induction.48,49 The presence of RNA transcripts in the vicinity of the 3′-bp deletion polymorphism shown by the ENCODE data (Figure 6) needs to be confirmed and explored further for their possible biologic functions. Additional investigations are needed to fully understand the molecular mechanisms and pathways involved in the modulation of HbF by this QTL.

The 3′-bp deletion, when present in people of either Chinese, European, or African descent, changes the normal DNA binding configuration of transcription factors and results in possible changes to the spatial orientation for DNA-protein binding and/or protein-protein interactions,50 which might account for the observed up-regulation of the enhancer-like activity. The consistent agreement of the association results across multiple populations, erythropoiesis-related transcription factor binding, phylogenetic conservation, and enhancer-like activity all suggest that the 3′-bp deletion polymorphism is probably the most significant functional variant within HMIP accounting for its modulation of HbF production.

Acknowledgments

The authors thank Ms Yvonne Chu and Ms Amanda Mok for their effort in subject recruitment in Hong Kong, Ms Stella Tsang for carrying out clinical laboratory testing, and all study subjects who agreed to participate in this research project.

This investigation was supported by National Institute of Diabetes and Digestive and Kidney Diseases (grant RO1 DK069646; D.H.K.C.) and National Heart, Lung, and Blood Institute (grant RO1 HL068970; M.H.S.).

Authorship

Contribution: J.J.F. and R.M.S. designed research, analyzed the data, and wrote the manuscript; Z.-y.C., H.-y.L., B.F.C., and C.T.B. undertook molecular testing and SNP genotyping, analyzed data, and edited the manuscript; S.Y.H., C.K.L., A.C.W.L., R.C.H.L., C.K.L., H.L.Y., J.C.C.S., E.S.K.M., L.C.C., and V.C. supervised subject recruitment and clinical laboratory testing, analyzed data, and edited the manuscript; and L.A.F., P.S., C.T.B., M.H.S., and D.H.K.C. conceived of and designed research, analyzed data, and wrote the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: David H. K. Chui, Evans 248, Department of Medicine, Boston University School of Medicine, 72 East Concord St, Boston, MA 02118; e-mail: david.chui@bmc.org.

References


A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression