Published in final edited form as:

Am J Hematol. 2016 June ; 91(6): E308–E311. doi:10.1002/ajh.24368.

Homozygosity for a Haplotype in the *HBG2-OR51B4* Region is Exclusive to Arab-Indian Haplotype Sickle Cell Anemia

Vinod Vathipadiekal¹, Abdulrahman Alsultan², Kristin Baltrusaitis³, John .J. Farrell¹, A.M. Al-Rubaish⁴, Fahad Al-Muhanna⁴, Z. Naserullah⁵, A. Alsuliman⁶, P.K. Patra⁷, Jacqueline N. Milton³, Lindsay A. Farrer¹, David H. K. Chui¹, Ameen K. Al-Ali⁸, Paola Sebastiani³, and Martin. H. Steinberg^{1,*}

¹ Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA

² Sickle Cell Disease Research Center and Department of Pediatrics, College of Medicine, King Saud University, Riyadh, Saudi Arabia

³ Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA

⁴ Department of Internal Medicine, College of Medicine, University of Dammam, Dammam, Kingdom of Saudi Arabia.

⁵ Alomran Scientific Chair for Hematological Diseases, King Faisal University & Departments of Pediatrics, Maternity & Child Hospital, Dammam, Kingdom of Saudi Arabia.

⁶ Alomran Scientific Chair, King Faisal University, King Fahd Hospital, Hafof, Al-Ahsa, Kingdom of Saudi Arabia.

⁷ Department of Biochemistry, Pt. J.N.M. Medical College, Raipur, Chattisgarh, India

⁸ Center for Research & Medical Consultation, University of Dammam, Dammam, Saudi Arabia

Keywords

Fetal hemoglobin; γ -globin gene expression; globin genes

Introduction

Homozygosity for rs334 (GAG-GTG, glu6val) or sickle cell anemia is common in Saudi Arabia. In the Eastern Province the sickle hemoglobin (HbS) gene is usually on the autochthonous Arab Indian (AI) β -globin gene (*HBB*) haplotype [1]. The higher fetal

* Corresponding Author: Martin H. Steinberg, 72 E. Concord St., Boston, MA 02118, 617-414-1020, FAX 617-414-1021; ; Email: mhsteinb@bu.edu.

Web Resources

Burrows-Wheeler Aligner, <http://bio-bwa.sourceforge.net/>

The Genome Analysis Toolkit, <https://www.broadinstitute.org/gatk/>

RegulomeDB Analysis, <http://www.regulomedb.org>

Regional plots of association results and recombination rates, <http://locuszoom.sph.umich.edu/locuszoom/>

Haplotype Analysis, <https://cran.r-project.org/web/packages/haplo.stats/index.html>

Supplemental data description

Supplemental Data include 5 figures and 3 tables, detailed methods for genotyping, ARMS assay, whole genome sequencing, and statistical analysis.

hemoglobin (HbF) level in the AI haplotype is often associated with milder disease compared with other HbS-associated haplotypes, especially in children [2, 3].

The XmnI G-A (C-T) polymorphism (rs7482144) is present in both the AI haplotype and the Senegal haplotype. Rs10128556, which is in linkage disequilibrium (LD) with rs7482144, could be the functional SNP of the Senegal haplotype [4]. The minor alleles of these SNPs also characterize the AI haplotype where adults had ~20% HbF compared with ~10% HbF in the Senegal haplotype. This suggests that elements linked to rs10128556 and rs7482144 in the AI but not the Senegal haplotype might affect HbF gene expression.

We compared the genome-wide distribution of SNPs in Saudi patients with the AI haplotype and Benin haplotype. Variants distinguishing these populations were limited to chromosome 11p15.5. Annotation of these SNPs narrowed the selection to 3 that we analyzed jointly using haplotype analysis. Homozygosity for a T/A/T haplotype of rs16912979 (in HS-4 of the LCR), rs7482144 and rs10128556 was exclusive to the AI haplotype. The AI haplotype might include a functionally important sub-haplotype that accounts for high HbF.

Methods

Patients

We recruited 245 sickle cell disease patients from the Eastern and Southwestern Provinces of Saudi Arabia for a genome-wide association study (GWAS). Forty-two AI (cohort 1) and 71 Benin cases were compared. Sixty-two additional 62 AI Saudi cases (cohort 2) and 41 Indian AI haplotype cases were studied by GWAS or targeted SNP analysis [5, 6]. Fourteen Saudi AI haplotype homozygotes had whole genome sequencing (WGS) including 7 AI homozygotes with high and 7 with low HbF, 3 Indian AI homozygotes, 3 African American Benin haplotype homozygotes selected because of unusually high HbF [7] and 1 African American Senegal haplotype homozygote. GWAS in 339 African American HbS homozygotes of diverse haplotypes examined the presence of SNPs found in the Saudi cohorts [8]. (**Table I**)

HBB haplotypes

The AI haplotype was ascertained directly by genotyping rs7482144, rs3834466 and rs549964658; other haplotypes were determined as described [6, 9, 10].

GWAS

Genotyping used Illumina SNP arrays [8].

Imputation

To capture all SNPs in the *HBB* gene cluster region, SHAPEIT and IMPUTE2 were used to impute non-genotyped SNPs using reference files from the 1000 Genome phase I database [11]. Only variants with a quality score threshold of 0.90, a minor allele frequency (MAF) difference >0.85 between the Saudi AI cohort1 and Saudi Benin haplotype cohorts, and SNPs annotated with RegulomeDB that were examined only in blood cells were included in our final analysis (**Fig. S1**)

Whole genome sequencing

Paired-end 100 bp reads were sequenced on the Illumina HiSeq 2000 platform to a depth of 40X per individual. (supplemental material) [12, 13].

Amplification refractory mutation system (ARMS)

ARMS was used to genotype rs16912979 in 41 AI haplotype Indian HbS homozygotes (supplementary material).

Epigenetics

Transcription factor binding data sets from the RegulomeDB of the ENCODE Consortium and Roadmap Epigenomics Project were searched for enhancer marks and transcription factor binding in blood cells [14].

Statistical analysis

Analysis was done in PLINK using logistic regression with additive coding of the SNPs [15]. Genomic control (GC) approach corrected the genomic inflation caused by sub-population stratification. QQ plots and genomic control estimates were generated using R (Fig. S2). Regional association plots were generated by the LocusZoom [16]. Haplotype analysis was conducted using the haplostats package in the R software as described [17]. (supplemental material)

Results

Directly genotyped SNPs

Two hundred-twenty three variants in chromosome 11p15.5 from positions 3.5 to 6.5 mb distinguished Saudi AI (cohort 1) and Benin cases (p-values $9.6E-07$ - $2.7E-45$). (Fig. 1A, Table SI). Thirteen SNPs were present in all Saudi AI haplotype but rare in the Saudi Benin haplotype (allele frequency < 0.05) (Table SII, Fig. S3). These results were replicated in the 62 AI haplotype cohort 2 cases and in 14 Saudi and 3 Indian cases genotyped by WGS. The 13 SNPs were not present in the 3 high HbF African American Benin and 1 Senegal haplotype samples and were rare or absent in 93 Senegal and 606 Benin haplotype chromosomes. A regional LD plot for rs16912979 is shown in Fig. S4. MAF of SNPs in *MYB*, *BCL11A* and *KLF1* were similar in AI and Benin haplotype cohorts.

Imputed SNPs

Ten SNPs were between *OR51B4* upstream of *HBB* and *OR52A5* downstream of *HBB*. Rs113622911 at coordinate 5269855 and rs59495893 at coordinate 5274126 were about 20 kb upstream of *HBB* in an area with H3K27Ac marks and POLR2A binding but were upstream of the canonical promoters of this gene. (Fig. 1B)

Epigenetic marks

Thirteen SNPs annotated by their putative functional role using the RegulomeDB database (Table SII). Rs16912979, rs4910743 and rs4601817 were within the *HBB* gene cluster LCR; rs16912979 was located in DNase1 HS-4. Rs6912979 lies with a region of H3K27Ac

marks and strong binding signals for POLR2A, GATA1, GATA2 and JUND (**Fig. 1B**); rs4601817 has weak binding signals for JUND and POLR2A; rs4910743 is in a region with H3K27Ac marks.

A unique AI haplotype

The epigenetic marks associated with rs16912979 and its presence in LCR HS-4 suggested this SNP as a component of a putative functional AI haplotype that included rs7482144 and rs10128556. Homozygosity for the minor allele of these 3 SNPs was limited to individuals with this haplotype. The T allele of rs16912979 was present in all 46 Bantu (CAR) haplotype chromosomes; however, these chromosomes had the T/T/C haplotype. (**Table SIII**)

Discussion

The Xmn1 G-A polymorphism or rs7482144, a marker of the Senegal *HBB* haplotype, was associated with increased *HBG2* expression [18]. Both rs7482144 and rs10128556 were also present in AI haplotype sickle cell anemia. These two SNPs, along with rs16912979, constituted a unique haplotype that included or was in LD with functional elements that might contribute to high HbF in AI haplotype. Homozygosity for this T/A/T haplotype distinguished the AI from all other *HBB* haplotypes. This suggested that maximum cis-acting modulation of HbF requires elements of both the AI and Senegal haplotype. Other variants exclusive to the AI haplotype within this region could also be the functional elements.

Transcription factor binding data and enhancer marks for some of the AI haplotype-associated SNPs suggested the presence of functional variants in this region and a haplotype effect on *HBG* expression [14]. We focused on rs16912979 because of its location in HS-4 of the LCR and strong binding signals for GATA1 and GATA2 and POLR2A in K562 cells. Recruitment of RNA polymerase II (Pol II) to the LCR, which is dependent on GATA1, is important for transcriptional activation of the downstream globin genes [19]. Relative concentrations of the GATA transcription factors play an important role in the regulation of *HBG* and *HBE* expression [20]. A binding signal for the large subunit of Pol II (POLR2A) is also present [21, 22]. Interrogation of all SNPs present at high frequency in the AI haplotype found many in regions with H3K27Ac marks in erythroid cells suggesting the open chromatin characteristic of an active enhancer [23].

The T/A/T sub-haplotype might mark a functional domain for the cis-acting regulation of *HBG2* expression [24]. HS-4, is required for high-level globin gene expression in definitive erythroid cells and contains a strictly conserved GATA1 binding site [25]. The distal LCR physically contacts the proximal globin gene promoters via chromatin looping that is developmentally and stage-specifically regulated [26]. DNA sequence motifs that are most conserved include GATA sequences in HS-2, HS-3 and HS-4, KLF1 binding sites in HS-2 and HS-3, and an E-box motif in HS-2 [27]. GATA1 is required for chromatin loop formation between hypersensitive sites and gene promoters [28]. LCR looping to globin gene promoters is facilitated by the LDB1/LMO2/GATA-1/TAL-1 erythroid specific protein complex (Ldb1 complex) [29-31]. TAL1 is a transcription factor that binds to regulatory

regions of many erythroid genes as part of a complex with GATA1, LMO2 and Ldb1; TAL1 binds at HS-4, HS-2 and HS-1 [29]. TAL1 overexpression increases its occupancy at HS-4 and HS-2 enhancing the expression of *HBG*. GATA1 and NF-E2 are both required for chromatin loop formation between the LCR and the active γ -globin genes in K562 cells [28]. RNA sequencing revealed a transcript from this region in human CD34⁺ cells and non-coding RNAs might play a role in modifying histones across the LCR and looping with *HBG2* promoter [32]. In adult erythroblasts that express *HBB*, forced chromosome looping using the self-association domain of the Ldb1 complex established strong interactions of the *HBG2* promoter with HS-2, -3 and -4, diminished interactions with the adult globin genes, and increased *HBG* expression to about 85% of total globin [33].

Variants cis to *HBB* that are not exclusive to the AI haplotype, in the context of homozygosity for the T/A/T haplotype might also account for further modulation of HbF. The T/A/T haplotype or a more extended haplotype of SNPs in LD, might be required for optimally functional looping of the LCR to the *HBG2* promoter and its robust transcription (Fig. S5)

These genetic association studies provide a rationale for functional studies of *HBG2* expression in wild-type and T/A/T haplotype erythroblasts and mechanistic studies like chromatin conformation capture experiments, to evaluate the role of chromatin looping as a mediator of the T/A/T haplotype effects on HbF.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funded in part by the University of Dammam, SP 11/2011, Office of Collaboration and Knowledge Exchange, University of Dammam, and R01 HL 068970, RC2 HL 101212, R01 87681, T32 HL007501 (VV) from the NHLBI, and T32GM074905 from the NIGMS Bethesda, MD

Charles Jahnke provided technical assistance with the use of the Boston University Medical Campus Linux Clusters for Genetic Analysis computing resource. Whole genome sequencing results are available on request from the University of Dammam and Boston University.

References

1. El-Hazmi MA, Al-Hazmi AM, Warsy AS. Sickle cell disease in Middle East Arab countries. *Indian J Med Res.* 2011; 134:597–610. [PubMed: 22199098]
2. Alsultan A, Alabdulaali MK, Griffin PJ, et al. Sickle cell disease in Saudi Arabia: the phenotype in adults with the Arab-Indian haplotype is not benign. *Br J Haematol.* 2014; 164:597–604. [PubMed: 24224700]
3. Perrine RP, Brown MJ, Clegg JB, et al. Benign sickle-cell anaemia. *Lancet.* 1972; 2:1163–1167. [PubMed: 4117591]
4. Galarneau G, Palmer CD, Sankaran VG, et al. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nature Genet.* 2010; 42:1049–1051. [PubMed: 21057501]
5. Sebastiani P, Farrell JJ, Alsultan A, et al. BCL11A enhancer haplotypes and fetal hemoglobin in sickle cell anemia. *Blood Cells Mol Dis.* 2015; 54:224–230. [PubMed: 25703683]

6. Ngo D, Bae H, Steinberg MH, et al. Fetal hemoglobin in sickle cell anemia: genetic studies of the Arab-Indian haplotype. *Blood Cells Mol Dis.* 2013; 51:22–26. [PubMed: 23465615]
7. Akinsheye I, Solovieff N, Ngo D, et al. Fetal hemoglobin in sickle cell anemia: molecular characterization of the unusually high fetal hemoglobin phenotype in African Americans. *Am J Hematol.* 2012; 87:217–219. [PubMed: 22139998]
8. Solovieff N, Milton JN, Hartley SW, et al. Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. *Blood.* 2010; 115:1815–1822. [PubMed: 20018918]
9. Steinberg MH, Hsu H, Nagel RL, et al. Gender and haplotype effects upon hematological manifestations of adult sickle cell anemia. *Am J Hematol.* 1995; 48:175–181. [PubMed: 7532353]
10. Alsultan A, Ngo D, Farrell JJ, et al. A functional promoter polymorphism of the δ -globin gene is a specific marker of the Arab-Indian haplotype. *Am J Hematol.* 2012; 87:824–826. [PubMed: 22641479]
11. Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.* 2007; 39:906–913. [PubMed: 17572673]
12. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* 2011; 43:491–498. [PubMed: 21478889]
13. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297–1303. [PubMed: 20644199]
14. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol.* 2010; 28:1045–1048. [PubMed: 20944595]
15. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–575. [PubMed: 17701901]
16. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010; 26:2336–2337. [PubMed: 20634204]
17. Lake SL, Lyon H, Tantisira K, et al. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered.* 2003; 55:56–65. [PubMed: 12890927]
18. Nagel RL, Fabry ME, Pagnier J, et al. Hematologically and genetically distinct forms of sickle cell anemia in Africa. The Senegal type and the Benin type. *N Engl J Med.* 1985; 312:880–884. [PubMed: 2579336]
19. Dean A. On a chromosome far, far away: LCRs and gene expression. *Trends Genet.* 2006; 22:38–45. [PubMed: 16309780]
20. Weiss MJ, Orkin SH. GATA transcription factors: key regulators of hematopoiesis. *Exptl Hematol.* 1995; 23:99–107.
21. Jonkers I, Lis JT. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol.* 2015; 16:167–177. [PubMed: 25693130]
22. Sainsbury S, Bernecky C, Cramer P. Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol.* 2015; 16:129–143. [PubMed: 25693126]
23. Creighton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci US A.* 2010; 107:21931–21936.
24. Dogan N, Wu W, Morrissey CS, et al. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin.* 2015; 8:16. [PubMed: 25984238]
25. Navas PA, Peterson KR, Li Q, et al. The 5'HS4 core element of the human beta-globin locus control region is required for high-level globin gene expression in definitive but not in primitive erythropoiesis. *J Mol Biol.* 2001; 312:17–26. [PubMed: 11545582]
26. Levings PP, Bungert J. The human beta-globin locus control region. *Eur J Biochem.* 2002; 269:1589–1599. [PubMed: 11895428]
27. Hardison R, Slightom JL, Gumucio DL, et al. Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. *Gene.* 1997; 205:73–94. [PubMed: 9461381]

28. Woon Kim Y, Kim S, Geun Kim C, et al. The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal gamma-globin genes. *Nucl Acids Res.* 2011; 39:6944–6955. [PubMed: 21609963]
29. Yun WJ, Kim YW, Kang Y, et al. The hematopoietic regulator TAL1 is required for chromatin looping between the beta-globin LCR and human gamma-globin genes to activate transcription. *Nucl Acids Res.* 2014; 42:4283–4293. [PubMed: 24470145]
30. Deng W, Lee J, Wang H, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell.* 2012; 149:1233–1244. [PubMed: 22682246]
31. Krivega I, Dale RK, Dean A. Role of LDB1 in the transition from chromatin looping to transcription activation. *Genes & Devel.* 2014; 28:1278–1290. [PubMed: 24874989]
32. Kim YW, Lee S, Yun J, et al. Chromatin looping and eRNA transcription precede the transcriptional activation of gene in the beta-globin locus. *Biosci Rep.* 2015:35.
33. Deng W, Rupon JW, Krivega I, et al. Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell.* 2014; 158:849–860. [PubMed: 25126789]

Figure 1A.

Manhattan plot from the GWAS comparing Saudi Eastern AI vs. Saudi Southwestern Benin haplotype patients. P-values ($-\log_{10} P$) of 599,131 SNPs after correction by Genomic Control is plotted against its physical chromosomal position. Odd chromosomes are in blue and even chromosomes in orange. Genome-wide significant variants separating these populations are clustered in chromosome 11p15.5.

Figure 1B. Epigenetic marks of AI haplotype-specific and other SNPs. Shown in the 1st track are the chromosomal locations of the genes of the *HBB* cluster including the upstream (*OR51B4*) and downstream (*OR52A5*) olfactory receptors genes. A blue box shows the position of hypersensitive site 4 (HS-4) of the LCR. The 2nd track in light blue shows marks for histone H3 acetyl lysine 27 (HeK27Ac) throughout this region. Track 3 (DNase) shows DNase1 hypersensitive sites and track 4 (TFBS) shows binding sites for the transcription factors that include GATA1, GATA2, POLR2A and JUND. Below this are the locations of the SNPs detected by GWAS (black arrowheads (**Table SII**)). SNPs shown in red arrowheads form a triallelic haplotype whose minor alleles are present only in the AI haplotype. Blue arrowheads show the positions of imputed SNPs that differ between Saudi AI and Saudi Benin haplotype patients (see text).

Table I

Patient cohorts

	Saudi AI Cohort 1 (n=42)	Saudi AI Cohort 2 (n=62)	Indian AI (n=41)	AA SEN.SEN (n=8)	AA SEN.het (n=75)	Saudi BEN.BEN (n=71)	AA BEN.BEN (n=232)	AA CAR.CAR (n=23)
HbF (%)	17.6±5.0	18.8±7.5	23.0±4.8	8.4±4.5	8.5±4.7	10.8±4.6	6.8±5.7	5.5±3.7
Age (yrs.)	26.4±11.1	23.9±9.6	14.6±4.7	32.1±12.3	19.4±13.9	18.6±11.0	22.0±16.3	20.0±16.8

	Saudi AI		Indian AI (n=3)	AA BEN.BEN (n=3)	AA SEN.SEN (n=1)
	↑HbF (n=7)	↓HbF (n=7)			
HbF (%)	23.5±2.6	8.2±1.3	26.0±4.5	19.8±0.4	16
Age (yrs.)	25.9±6.8	34.1±10.3	22.7±5.5	26.0±3.6	5.9

A. Saudi AI (cohort 1) and Saudi Benin haplotype patients were selected by their region of origin in the Eastern and Southwestern Provinces of Saudi Arabia. Saudi AI (cohort 2) was an independent patient group. Indian AI patients were from Raipur, Chattisgarh, India. Other cohorts include African Americans (AA) homozygous for Benin, Bantu (CAR) and Senegal haplotypes and Senegal haplotype compound heterozygotes with Benin and CAR haplotypes [18]. AI-AI homozygotes; SEN.SEN-Senegal homozygotes; SEN.het-Senegal compound heterozygotes; BEN.BEN-Benin homozygotes; CAR.CAR-CAR homozygotes. Values are mean ± SD. **B.** In addition to the 14 Saudi AI haplotype patients, African American samples chosen for whole genome sequencing included 3 Benin and 1 Senegal haplotype homozygotes. The Benin cases were selected for WGS because of their unusually high HbF, as reported previously [7].