Published in final edited form as:
Am J Hematol. 2010 January ; 85(1): 29–35. doi:10.1002/ajh.21572.

Genetic modifiers of the severity of sickle cell anemia identified through a genome-wide association study

Paola Sebastiani[1,*], **Nadia Solovieff**[1], **Stephen W. Hartley**[1], **Jacqueline N. Milton**[1], **Alberto Riva**[2], **Daniel A. Dworkis**[3], **Efthymia Melista**[3], **Elizabeth S. Klings**[3], **Melanie E. Garrett**[4], **Marilyn J. Telen**[4], **Allison Ashley-Koch**[4], **Clinton T. Baldwin**[5], and **Martin H. Steinberg**[3]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts

[2]Department of Molecular Genetics, University of Florida, Gainesville, Florida

[3]Department of Medicine, Boston University School of Medicine, Boston, Massachusetts

[4]Department of Medicine, Duke University School of Medicine, Durham, North Carolina

[5]Center for Human Genetics, Boston University School of Medicine, Boston, Massachusetts

Abstract

We conducted a genome-wide association study (GWAS) to discover single nucleotide polymorphisms (SNPs) associated with the severity of sickle cell anemia in 1,265 patients with either “severe” or “mild” disease based on a network model of disease severity. We analyzed data using single SNP analysis and a novel SNP set enrichment analysis (SSEA) developed to discover clusters of associated SNPs. Single SNP analysis discovered 40 SNPs that were strongly associated with sickle cell severity (odds for association >1,000); of the 32 that we could analyze in an independent set of 163 patients, five replicated, eight showed consistent effects although failed to reach statistical significance, whereas 19 did not show any convincing association. Among the replicated associations are SNPs in *KCNK6* a K^+ channel gene. SSEA identified 27 genes with a strong enrichment of significant SNPs ($P < 10^{-6}$); 20 were replicated with varying degrees of confidence. Among the novel findings identified by SSEA is the telomere length regulator gene *TNKS*. These studies are the first to use GWAS to understand the genetic diversity that accounts the phenotypic heterogeneity sickle cell anemia as estimated by an integrated model of severity. Additional validation, resequencing, and functional studies to understand the biology and reveal mechanisms by which candidate genes might have their effects are the future goals of this work.

Introduction

Phenotypic heterogeneity is characteristic of sickle cell anemia, a Mendelian disorder caused by homozygosity for the sickle *HBB* gene (glu6val). Among patients, different rates of hemolysis/vasculopathy and viscosity/vasoocclusion-related complications are typical [1–3], and account for a substantial reduction in life expectancy: In 1994, the median life expectancy

[*]Correspondence to: Department of Biostatistics, Boston University School of Public Health, Boston, MA. sebas@bu.edu .

Author Contributions PS designed the study, conducted the statistical analysis, contributed to the interpretation of results, and wrote the manuscript; NT, SWH, JNM, AR, AAK, and MEG contributed to the statistical analysis; DAW contributed to the interpretation of results; ESK and MJT provided data for replication, contributed to the interpretation of results, and the writing of the manuscript; CTB contributed to the execution of the study, the interpretation of the results, and the writing of the manuscript; MHS designed and conducted the study and wrote the manuscript.

Additional Supporting Information may be found in the online version of this article

Conflict of interest: Nothing to report

for men and women with sickle cell anemia was 42 and 48 years, respectively [4]. Despite many advances in care, the annual mortality still approaches 4%. To integrate individual disease complications into a comprehensive measure of severity, we developed a model of the associations among clinical and laboratory variables that scored disease severity as the risk of death within 5 years [5]. This network was developed using data obtained from more than 3,400 subjects from the Cooperative Study of Sickle Cell Disease (CSSCD) [6], and its accuracy was validated in two unrelated sets of sickle cell patients. Recently, the network was also validated in a small European cohort of patients with sickle cell anemia [7].

This model did not include the genetic components that likely determine the substantial between-subject variability of sickle cell anemia complications. Preliminary work based on candidate gene analysis suggested that several genes might modulate the overall severity of the disease but was limited to the choice of candidate genes [8]. We report here the results of a genome-wide association study (GWAS) of the severity of sickle cell anemia in 1,265 patients from the CSSCD based on the Illumina Human610-Quad SNP array. Using traditional single SNP analysis, we identified 40 SNPs strongly associated with disease severity (odds for association >1,000) and, of the 32 that we could analyze in an independent validation set of 163 patients, five replicated, eight showed consistent effects although failed to reach statistical significance, and 19 were not significant.

The small number of replicated associations that were significant is a limitation of the standard approach to GWAS that usually requires a very large sample size to detect associations with so-called genome-wide significance (P values < 10^{-8}) to reduce chances of false positive findings [9]. It is acknowledged that many more associations remain to be discovered from GWAS data [10], especially in rare diseases where sample sizes are constrained. Therefore, we developed a novel SNP set enrichment analysis (SSEA) to mine the GWAS data further. SSEA searches for regions of the genome that are enriched for sets of significantly associated SNPs and computes the probability of detecting by chance the number of SNPs associated with a phenotype in a window of a flexible number of SNPs, or in a gene. SSEA identified 18 genes with a strong enrichment of significant SNPs in the discovery set ($P < 2 \times 10^{-6}$) and 11 were replicated with varying degrees of confidence. Functional category enrichment analysis showed an enrichment of genes involved in heme binding.

Methods

Subjects

The CSSCD remains the largest study of sickle cell disease yet done in the United States [6]. Although it has been more than 30 years since the first patient was enrolled, it is unlikely to be replicated soon. In the CSSCD, phenotype and laboratory data were collected according to protocol, quality checked, and biological samples obtained from patients with sickle cell anemia [11]. We used 1,265 sickle cell anemia patients, all African-Americans, from the CSSCD as our primary discovery set for the GWAS, using stored DNA for genotyping. Our secondary validation data sets were from much smaller contemporaneous sickle cell anemia studies and included African-Americans subjects being screened prospectively for the presence of pulmonary hypertension from the Duke University Pulmonary Hypertension Study [12] and from Boston Medical Center (BMC) [13]. Some of the latter patients were previously included in the validation of the score of sickle cell anemia severity [5]. In both the Duke and Boston studies, patients were at least 18 years old. In the CSSCD, patients were recruited regardless of symptoms in an attempt to capture the spectrum of disease severity. The Duke and BMC samples were clinic-based, and therefore, biased toward symptomatic disease. These studies were approved by the Institutional Review Boards of the participating institutions. Summary characteristics of the populations are described in Table I.

Phenotype

We used the Bayesian network described in [5] and available from http://www.bu.edu/sicklecell/downloads/Projects/ to compute the severity of disease of patients of the discovery and validation sets. The network uses 25 clinical and laboratory variables to assess the severity of disease by the risk of death within 5 years and its sensitivity and specificity were originally evaluated in two unrelated patient data sets. The network model includes well-known markers of disease severity such as LDH and systolic blood pressure that correlate with hemolytic anemia, and other complications of the disease such as stroke, painful episodes, and priapism that are known markers of severity. The quantitative score was used to generate a group of 1,088 patients with “mild disease” (score < 0.4) and 177 patients with “severe disease” (score ≥0.6 for ages 40 and younger, and score >0.8 for ages 40 and older) that were used as controls and cases for the GWAS. This grouping was based on the observation that the severity score has a U shape that changes in age groups (Fig. 1). The lower value of 0.4 appears to capture groups of mild disease patients for all age groups, and a score of 0.6 or higher is sufficient to characterize severe patients aged 40 or younger. For patients aged 40 and older, we restricted the definition of severe cases to a score of 0.8 or higher to reduce the risk of misclassification. The severity of disease in the 163 subjects of the validation set was scored in analogous way and lead to 95 mild cases (score < 0.4) and 68 severe cases (score > 0.6).

Genotype

We genotyped the DNA samples of the CSSCD and Duke subjects with the Illumina Human610-Quad SNP array that includes more than 600,000 SNPs and covers ~60% of the genome of the HapMap Yoruban ($r^2 > 0.8$) and 18,000 known genes. We genotyped the DNA samples from the BMC patients in an earlier phase of the study, with the Illumina HumanCNV370-duo bead chip that comprises ~350,000 SNPs common to the 610 array. Genomic DNA (0.5–1 μg depending on array) was analyzed on the Illumina arrays using the standard Illumina protocol and BeadStudio software was used for genotype calling using predetermined clustering provided by Illumina.

Quality control

We excluded samples with gender inconsistencies and call rates <93%. We conducted identity by state analysis using the software PLINK [14] to confirm known familial relationships and to identify hidden familial relationships. Repeat genotyping of duplicate samples received from the CSSCD as well as repeat genotyping of the same tube of DNA was used to evaluate sample misidentification (estimated to be <5%). All suspect samples were excluded from the study (~10%). We conducted genome-wide principal component analysis of the subjects included in the discovery and validation sets with the program EIGENSOFT [15]. The scatter plot of the first two principal components is displayed in Fig. 2.

Single SNP analysis

We examined general, allelic, dominant, and recessive associations using Bayesian tests and scored the evidence of association of each model by its posterior probability. The details of the calculations are as described in Balding [16]. and Sebastiani et al [17]. and we assumed uniform probability on competitive models, so that the posterior odds of each model is equivalent to the Bayes factor (BF). To address the issue of multiple testing, we conducted extensive simulations to compute the expected number of false positive associations for different thresholds of the BF. The simulations showed that the false positive rate of the Bayesian decision rule changes with the allele frequency and suggested using a BF >1,000 to reduce the number of false positive associations to less than 1 in 10,000 independent tests. This procedure is also described in [18]. Forty SNPs had a model of association with a BF >1,000 (Supporting

Table 1). Of these 40 SNPs, only 32 had an estimated MAF in the validation set of at least 15% and could be included in the validation analysis. A SNP association was considered validated if the same genetic model showed significant associations in both the discovery and validation set, and the genetic effects were in the same direction. However, the threshold for significance in the validation set was reduced to BF >1 given that replication was conducted for 32 SNPs only. A SNP association was considered consistent if the genetic effect for the same genetic model was in the same direction in both the discovery and validation set, but statistical significance was not reached in the validation set. The results are in Tables II and III. For comparisons, allele frequencies in the HapMap CEPH and Yoruban are also reported.

SSEA

Traditional one-SNP-at-a-time methods are underpowered to discover those myriad variants that explain minor effects upon common complex phenotypes or diseases. Therefore, for better mining of the data, we developed a SNP set enrichment analysis or SSEA. SSEA computes the probability of detecting by chance the number of SNPs associated with a phenotype within a set of SNPs, for a given threshold of significance. The smaller this probability, the stronger the evidence that the SNPs set is globally associated with the phenotype. The probability is computed using the hypergeometric distribution

$$\frac{\binom{m_g}{k_g}\binom{N-m_g}{n-k_g}}{\binom{N}{n}}$$

where N is the total number of SNPs analyzed, n is the total number of SNPs with BF > threshold, m_g is the number of SNPs in the gth set, and k_g is the number of SNPs in the gth set with a BF > threshold. We used a “gene-centric” definition of SNP sets that were defined as the set of SNPs in each gene transcript, including SNPs within 10 kb from the transcripts. We used a BF >3 to select significant SNPs in genes based on the following consideration: The mean number of SNPs per gene in the Illumina 610 is 25, and because the decision rule to accept an association as significant if the BF >3 has a false positive rate of about 2%, the expected number of false positive single SNP association per gene is on average 5%. This analysis identifies 91 genes with a SSEA P <0.001. The genes with significant SSEA scores were examined in the validation set and 20 of them had one or more SNPs significantly associated with sickle cell anemia severity. We report the 11 genes with a SSEA $P < 10^{-6}$ in the discovery set and SSEA P <0.2 in the validation set in Table IV. The threshold 10^{-6} limits the expected number of false positive findings to less than 5% in the discovery step, whereas the threshold of 0.2 for replication limits the expected number of falsely replicated genes to 4 and a false discovery rate of 4/11 = 0.36. Note that SSEA searches for clusters of significant SNPs that are individually mildly significant. Therefore, this analysis is complementary to the single SNP analysis.

The list of 91 genes with a SSEA P <0.001, equivalent to an expected false discovery rate of 10%, was functionally annotated using EASE [19] and the results are in Table V.

Results

Table I summarizes the patients’ characteristics of the discovery and validation sets. As expected, patients with severe disease in the primary discovery study are on average older than milder cases, have a substantially higher prevalence of osteonecrosis (AVN), leg ulceration, stroke, and sepsis but essentially no difference in frequency of pain episodes and priapism and

HbF concentration. The clinical characteristics of the contemporary validation set reflect the overall older age of this clinic population, the greater use of transfusion, and for LDH, different test characteristics and therefore, normal ranges in the discovery and validation samples. While widely separated in time from the CSSCD patients, the patients in the validation set had similar degrees of genetic admixture, and our analysis of the level of genetic diversity in the patients included in this study using principal component analysis [20] did not show differences in ancestry (Fig. 2). Full details of the principal component analysis are described in the manuscript [21].

Figure 3 displays the Manhattan plot with the results of the GWAS, using single SNP analysis. Forty SNPs reach our genome-wide significance threshold in at least one model of association (Supporting Table 1). Table II reports the eight SNPs that reach genome-wide significance in the discovery set and show consistent effects in the validation set. The results are shown for different models of genetic inheritance, and when more than one model reached genome-wide significance, the model with strongest significance and replication is reported. Table III reports the five SNPs that are associated with sickle cell anemia severity in the primary study and replicated in the secondary study. Besides the SNPs in *KCNK6* and *OTUD3*, the other three are in intergenic regions with no known functional role.

Table IV shows the list of 11 genes that have a significant SSEA probability score $<10^{-6}$. *TNKS* is the gene with strongest SSEA score in the discovery set, with 14 of 35 SNPs that were associated with a BF >3. The linkage disequilibrium heatmap in Fig. 4 shows the physical positions of the associated SNPs and their linkage disequilibrium map. The gene spans more than 240 Kb and the patterns of associations identify two blocks that would need to be followed by fine mapping or sequencing for SNP discovery.

Table V shows the top functional categories that were enriched of genes with SSEA score <0.001. The full list of 92 genes included in this analysis is in the Supporting Table 2. Noticeable gene categories are the set of three genes *CYP2C18* (SSEA score 0.0003); *CYP2C9* (SSEA score 5×10^{-5}); and *CYP4F8* (SSEA score 10^{-5}) that are members of the cytochrome P450 superfamily of enzymes. The full set of results is available also at http://www.bu.edu/sicklecell/projects/.

Discussion

Phenotypic heterogeneity is characteristic of sickle cell anemia and is partially genetically influenced [22]. GWAS are an unbiased approach toward discovering potential genetic modulators by seeking inherited polymorphisms associated with phenotypic traits. Single laboratory measurements, like HbF and pulmonary tricuspid regurgitant velocity on echocardiography, or single clinical events like stroke, priapism, osteonecrosis, pain, and leg ulceration, have been used as subphenotypes in genetic association studies [12,23–30]. We chose to use an integrated estimate of the severity of disease as a phenotype to better model global disease severity.

Our studies support the plausible notion that the actions of multiple genes determine the overall severity of sickle cell anemia. The modulatory effects of HbF and α thalassemia have been amply demonstrated (for a review see [22]) and genes modulating HbF concentration have been associated with the rate of acute painful episodes [31]. GWAS using a novel phenotype that reflects the overall severity of disease defined by death suggests an association with genes whose properties are pertinent to the pathophysiology of sickle cell anemia.

KCNK6 is a member of a superfamily of K^+ channel proteins. Widely expressed in CD34+ cells induced toward erythroid differentiation and in endothelium across different vascular beds, we found its expression upregulated when human pulmonary artery endothelial cells were exposed to sickle cell plasma [32]. The vascular endothelium plays a major role in vasoocclusive processes in sickle cell disease via its interactions with sickle erythrocytes, leukocytes, and platelets making it an important potential target for genetic modulation [33]. Also, sickle erythrocyte cation content is a critical determinant of HbS polymerization tendency that leads to cell damage and hemolysis. Whether or not this gene is expressed in reticulocytes or mature erythrocytes is unknown. An analysis of the reticulocyte transcriptome did not detect its expression among the 120 most expressed genes. However, cation transport channel genes like the Gardos channel (*KCCN4*) or KCl cotransporter (*KCC1*, *KCC3*, *KCC4*) channels that are known to be active in sickle reticulocytes were also not represented among the highly expressed genes of the normal reticulocyte transcriptome [32–36].

Tankyrase1, the protein product of *TNKS*, is a poly(ADP-ribose) polymerase that localizes to telomeres. Nuclear overexpression of *TNKS* leads to loss of telomere elongation, suggesting a role as a positive regulator of telomere length. Conversely, reducing *TNKS* expression using RNAi produces telomere shortening [37]. As sickle cell anemia, as well as many other vascular diseases, are characterized by endothelial dysfunction, of interest is the role that telomerases might play in this process [38,39]. Endothelial senescence is modulated by telomerase activity, and the pro-oxidant environment in sickle cell anemia might also lead to telomere shortening. Additionally, hydroxyurea, an important drug therapy for sickle cell anemia that prolongs survival, affects telomere replication and maintenance through a mechanism that may involve the direct modification of *TRF2*, a member of the shelertrin complex modulated by *TNKS*. *TRF2* can protect *TRF1* from the effects of tankyrase1 [40].

Several genes associated with severity by SSEA have unknown functions. Two genes identified by SSEA have interesting functions: The protein kinase *MAP3K13* may have a role in the JNK pathway that is activated by inflammation [41]. The complement component 8 (*C8A*) is one of the components of the membrane attack complex, but is not required for inducing hemolysis or for killing gram-negative bacteria[42,43]. Sixteen SNPs in this gene are in the 610 array, and eight are significantly associated with disease severity, including the functional SNP rs652785 (CAA → AAA position 414, Q93K). Patients with the AA genotype have 3.2 times the odds for severe disease compared to patients with the AC or CC genotype.

Although our study is a first step using GWAS to help understand the genetic diversity that accounts for the phenotypic heterogeneity of sickle cell anemia, the small size of the validation set limits our ability to validate the findings of the discovery set. Sickle cell anemia is a rare disease in developed countries where extensive phenotype data and biological samples are most likely to be available for genetic studies, thereby constraining the sample size needed to achieve a traditional genome-wide level of significance for individual SNPs with small effect sizes. This also compromises the ability to assemble replication and validation populations of similar ethnic composition. Our results will require additional validation, resequencing of promising leads, and functional studies to understand the biology and reveal mechanisms by which candidate genes might have their effects. However, they form a base for other investigators to use in their work on this subject. Any therapeutic use of our observations are likely to be years away. Nevertheless, when we more fully understand the genetic differences among patients and the association with phenotypes of disease this might be prognostically useful.[27,29]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Contract grant sponsor: NIH/NHLBI; Contract grant numbers: R01 HL87681, R01 HL 068970.

References

1. Kato GJ, Gladwin MT, Steinberg MH. Deconstructing sickle cell disease: Reappraisal of the role of hemolysis in the development of clinical subphenotypes. Blood Rev 2007;21:37–47. [PubMed: 17084951]

2. Serjeant GR, Higgs DR, Hambleton IR. Elderly survivors with homozygous sickle cell disease. N Engl J Med 2007;356:642–643. [PubMed: 17287491]

3. Taylor JG VI, Nolan VG, Mendelsohn L, et al. Chronic hyper-hemolysis in sickle cell anemia: Association of vascular complications and mortality with less frequent vasoocclusive pain. PLoS ONE 2008;3:e2095. [PubMed: 18461136]

4. Platt OS, Brambilla DJ, Rosse WF, et al. Mortality in sickle cell disease. Life expectancy and risk factors for early death. N Engl J Med 1994;330:1639–1644. [PubMed: 7993409]

5. Sebastiani P, Nolan VG, Baldwin CT, et al. A network model to predict the risk of death in sickle cell disease. Blood 2007;110:2727–2735. [PubMed: 17600133]

6. Gaston M, Rosse WF. The cooperative study of sickle cell disease: Review of study design and objectives. Am J Pediatr Hematol Oncol 1982;4:197–201. [PubMed: 7114401]

7. Anoop P, Bevan DH, Chakrabarti S. Usefulness and limitations of Bayesian network model as a mortality risk assessment tool in sickle cell anemia. Am J Hematol 2009;84:312–313. [PubMed: 19338040]

8. Sebastiani P, Wang L, Perls TT, et al. A repertoire of genes modifying the risk of death in sickle cell anemia. Blood (ASH Annu Meeting Abstracts) 2007;110:150.

9. Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008;299:1335–1344. [PubMed: 18349094]

10. Donnelly P. Progress and challenges in genome-wide association studies in humans. Nature 2008;456:728–731. [PubMed: 19079049]

11. West MS, Wethers D, Smith J, Steinberg M. Laboratory profile of sickle cell disease: A cross-sectional analysis. The Cooperative Study of Sickle Cell Disease. J Clin Epidemiol 1992;45:893–909. [PubMed: 1624972]

12. Ashley-Koch AE, Elliott L, Kail ME, et al. Identification of genetic polymorphisms associated with risk for pulmonary hypertension in sickle cell disease. Blood 2008;111:5721–5726. [PubMed: 18187665]

13. Klings ES. Pulmonary hypertension of sickle cell disease: More than just another lung disease. Am J Hematol 2008;83:4–5. [PubMed: 17924550]

14. 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]

15. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909. [PubMed: 16862161]

16. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;7:781–791. [PubMed: 16983374]

17. Sebastiani P, Zhao Z, Abad-Grau MM, et al. A hierarchical and modular approach to the discovery of robust associations in genome-wide association studies from pooled DNA samples. BMC Genet 2008;9:6. [PubMed: 18194558]

18. Sebastiani P, Timofeev N, Dworkis DA, et al. Genome-wide association studies and the genetic dissection of complex traits. Am J Hematol 2009;84:504–515. [PubMed: 19569043]

19. Hosack DA, Dennis G, Sherman BT, et al. Identifying biological themes within lists of genes with EASE. Genome Biol 2003;4:R70. [PubMed: 14519205]

20. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2:e190. [PubMed: 17194218]

21. Timofeev N, Sebastiani P, Gladwin M, et al. 3rd Annual Sickle Disease Research and Educational Symposium and Annual National Sickle Cell Disease Scientific Meeting, Ft Lauderdale. Am J Hematol 2009;84:E1–E235.

22. Steinberg MH. Genetic etiologies for phenotypic diversity in sickle cell anemia. ScientificWorldJournal 2009;9:46–67. [PubMed: 19151898]

23. Hoppe C, Klitz W, Noble J, et al. Distinct HLA associations by stroke subtype in children with sickle cell anemia. Blood 2003;101:2865–2869. [PubMed: 12517810]

24. Hoppe C, Klitz W, Cheng S, et al. Gene interactions and stroke risk in children with sickle cell anemia. Blood 2004;103:2391–2396. [PubMed: 14615367]

25. Baldwin C, Nolan VG, Wyszynski DF, et al. Association of klotho, bone morphogenic protein 6, and annexin A2 polymorphisms with sickle cell osteonecrosis. Blood 2005;106:372–375. [PubMed: 15784727]

26. Nolan VG, Baldwin C, Ma Q, et al. Association of single nucleotide polymorphisms in klotho with priapism in sickle cell anaemia. Br J Haematol 2005;128:266–272. [PubMed: 15638863]

27. Sebastiani P, Ramoni MF, Nolan V, et al. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet 2005;37:435–440. [PubMed: 15778708]

28. Nolan VG, Adewoye A, Baldwin C, et al. Sickle cell leg ulcers: Associations with haemolysis and SNPs in klotho, TEK and genes of the TGF-β/BMP pathway. Br J Haematol 2006;133:570–578. [PubMed: 16681647]

29. Uda M, Galanello R, Sanna S, et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. Proc Natl Acad Sci USA 2008;105:1620–1625. [PubMed: 18245381]

30. Sedgewick AE, Timofeev N, Sebastiani P, et al. BCL11A is a major HbF quantitative trait locus in three different populations with beta-hemoglobinopathies. Blood Cells Mol Dis 2008;41:255–258. [PubMed: 18691915]

31. Lettre G, Sankaran VG, Bezerra MA, et al. DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. Proc Natl Acad Sci USA 2008;105:11869–11874. [PubMed: 18667698]

32. Klings ES, Safaya S, Adewoye AH, et al. Differential gene expression in pulmonary artery endothelial cells exposed to sickle cell plasma. Physiol Genomics 2005;21:293–298. [PubMed: 15741505]

33. Hebbel RP, Osarogiagbon R, Kaul D. The endothelial biology of sickle cell disease: Inflammation and a chronic vasculopathy. Microcirculation 2004;11:129–151. [PubMed: 15280088]

34. Brugnara, C. Red cell membrane in sickle cell disease. In: Steinberg, MH.; Forget, BG.; Higgs, DR.; Nagel, RL., editors. Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management. Cambridge University Press; Cambridge: 2001. p. 550-576.

35. Rajashekhar G, Grow M, Willuweit A, et al. Divergent and convergent effects on gene expression and function in acute versus chronic endothelial activation. Physiol Genomics 2007;31:104–113. [PubMed: 17566077]

36. Goh SH, Josleyn M, Lee YT, et al. The human reticulocyte transcriptome. Physiol Genomics 2007;30:172–178. [PubMed: 17405831]

37. Donigian JR, De Lange T. The role of the poly(ADP-ribose) polymerase tankyrase1 in telomere length control by the TRF1 component of the shelterin complex. J Biol Chem 2007;282:22662–22667. [PubMed: 17561506]

38. Erusalimsky JD. Vascular endothelial senescence: From mechanisms to pathophysiology. J Appl Physiol 2009;106:326–332. [PubMed: 19036896]

39. Erusalimsky JD, Skene C. Mechanisms of endothelial senescence. Exp Physiol 2009;94:299–304. [PubMed: 18931048]

40. Snyder AR, Zhou J, Deng Z, Lieberman PM. Therapeutic doses of hydroxyurea cause telomere dysfunction and reduce TRF2 binding to telomeres. Cancer Biol Ther 2009;8:1136–1145. [PubMed: 19363303]

41. Ikeda A, Hasegawa K, Masaki M, et al. Mixed lineage kinase LZK forms a functional signaling complex with JIP-1, a scaffold protein of the c-Jun NH(2)-terminal kinase pathway. J Biochem 2001;130:773–781. [PubMed: 11726277]

42. Parker CL, Sodetz JM. Role of the human C8 subunits in complement-mediated bacterial killing: Evidence that C8 gamma is not essential. Mol Immunol 2002;39:453–458. [PubMed: 12413696]

43. Esser AF. The membrane attack complex of complement. Assembly, structure and cytotoxic activity. Toxicology 1994;87:229–247. [PubMed: 8160186]

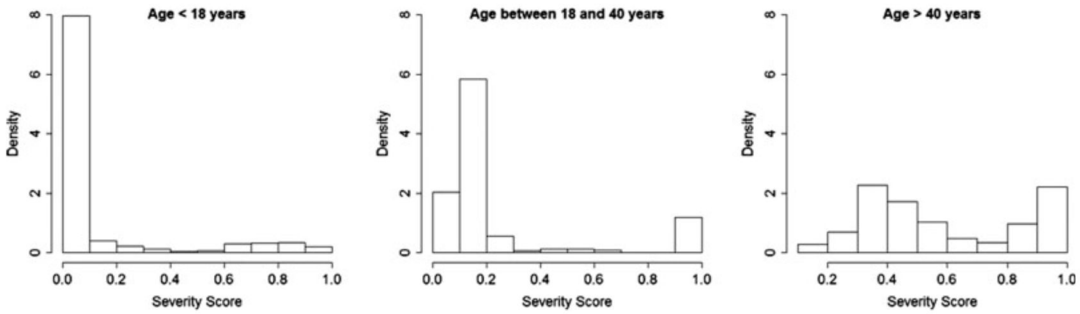

Figure 1.

Distribution of the severity score in patients of the CSSCD. The three histograms show the different distribution of disease severity score in three age groups (655 patients, age < 18; 504 ages between 18 and 40 years, and 150 patients, age 40 and older). Note that each y-axis reports the density, so that the area of each bar is the relative frequency of patients with score within the limits in the x-axis.

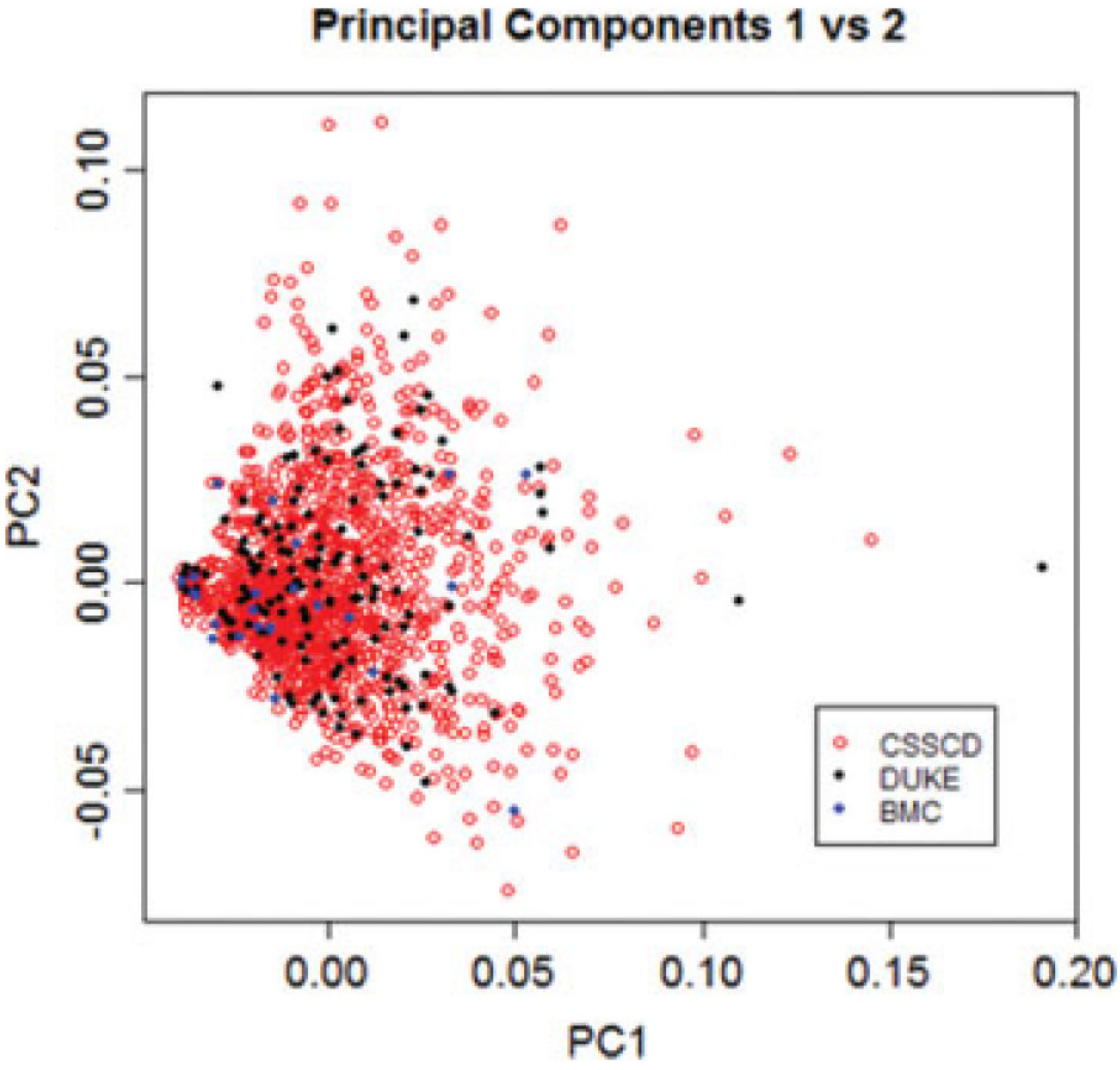

Figure 2.

Display of the principal components 1 and 2 that capture the largest amount of genetic variability in the subjects of the discovery and validation sets. The total overlapping of principal components 1 (PC1, *x*-axis) and 2 (PC2, *y*-axis) in subjects of the discovery set (CSSCD) and validation set (BMC and Duke) shows that the subjects in the two sets have similar genetic diversity. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

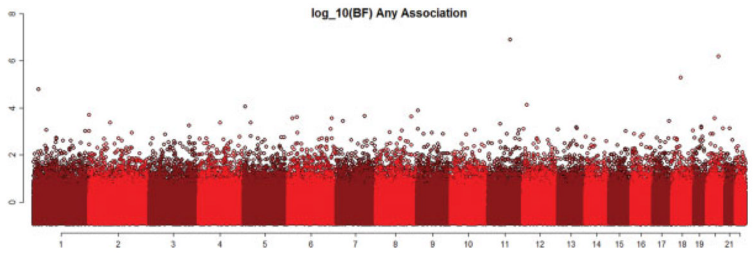

Figure 3.

Manhattan plot displaying the log10 (Bayes Factor) for the GWAS of severity of sickle cell disease. We tested the association of each SNP with severity of sickle cell disease using general, allelic, dominant, and recessive models. The *x*-axis reports the physical positions of SNPs in chromosomes 1–22, and the *y*-axis reports the maximum log10 Bayes factor observed for each SNP. Genome-wide significant is met with a Bayes factor >1,000 (log10 Bayes factor >3). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

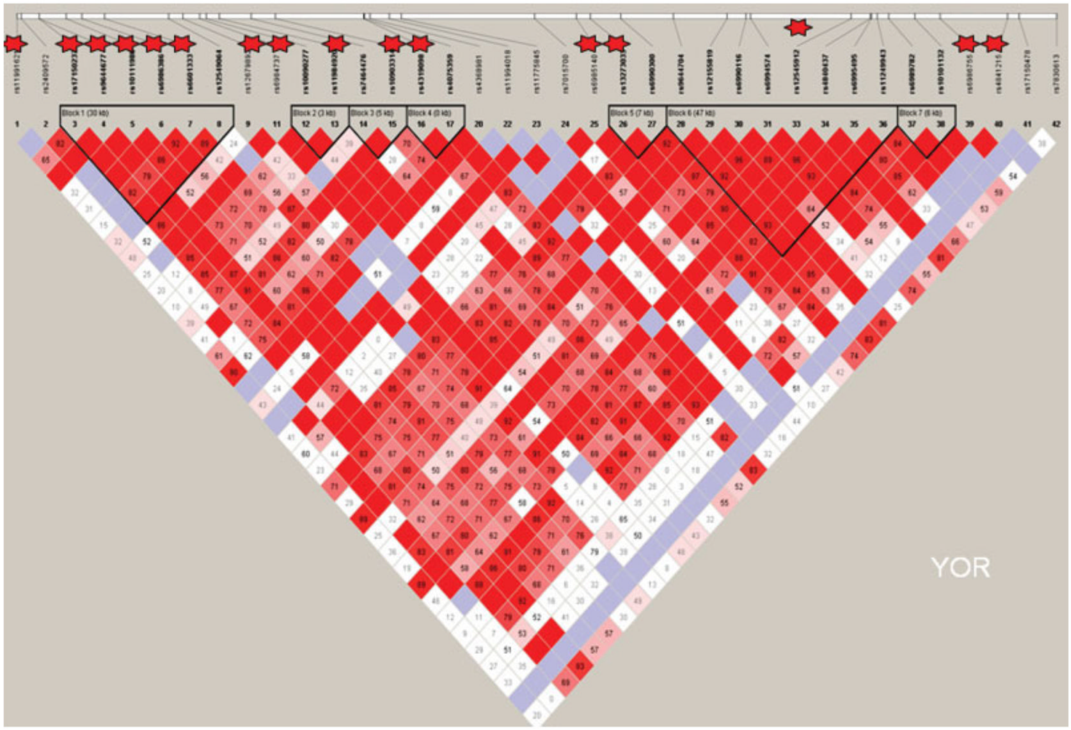

Figure 4.

LD heatmap of the gene *TNKS*, using the HapMap Yoruban. Red stars denote the SNPs that are associated with SCA severity. The heatmap displays the estimate of LD using *D*’ and was produced using the software Haploview and the data from HapMap Yorubans. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

TABLE I

Summary of Patients Characteristics

	CSSCD				Duke + BMC			
	Severe patients		**Mild patients**		**Severe patients**		**Mild patients**	
	***n* = 177**		***n* = 1,088**		***n* = 68**		***n* = 95**	
	Mean	**StDev**	**Mean**	**StDev**	**Mean**	**StDev**	**Mean**	**StDev**
Age[a,b]	23.97	17.33	18.43	11.73	43.55	11.16	31.80	12.27
ALT level, Sgpt, U/L[b]	38.90	59.76	36.46	51.96	32.58	21.24	24.55	14.72
AST level, Sgot, U/L[a,b]	54.61	29.50	46.97	20.90	53.13	31.85	39.63	17.51
Bilirubin level, mg/dL	3.03	1.78	3.13	1.89	2.92	3.28	3.20	3.14
BUN level, mg/dL[a,b]	10.59	7.60	8.77	4.80	14.28	13.81	10.07	6.83
Creatinine level, mg/dL[b]	0.87	1.40	0.61	0.42	1.16	1.02	0.98	1.42
Hb level, g/dL	8.35	1.37	8.44	1.27	8.34	1.63	8.61	1.63
% HbF	6.48	5.64	6.60	5.65	7.44	6.85	6.28	3.74
LDH level, U/L	495.84	193.16	496.89	203.38	1162.06	713.73	1038.72	696.21
MCV, fL[a]	90.47	7.86	89.34	8.03	89.87	11.55	89.88	9.67
Platelet count, 1000/L[a,b]	405.16	118.89	434.04	112.98	408.04	152.79	470.20	157.42
Reticulocyte count, % rRC	11.73	4.94	11.39	5.32	11.81	6.81	11.73	7.36
Sys BP, mmHg[a,b]	107.73	13.54	105.03	10.55	132.00	19.40	118.04	12.07
WBC count, 1000/L[a]	12.62	2.85	11.92	2.79	11.57	3.57	11.87	3.59

	Proportion	**Proportion**	**Proportion**	**Proportion**
ACS[a,b]	0.79	0.68	0.82	0.74
AVN	0.41	0.35	0.48	0.34
Blood transfusion[a]	0.23	0.14	0.46	0.46
Death[a,b]	0.36	0.01	0.14	0.08
Leg ulceration[a,b]	0.23	0.12	0.22	0.19
Pain	0.92	0.89	0.77	0.71
Priapism	0.18	0.18	0.58	0.18
Sepsis[a,b]	0.93	0.01	0.17	0.01
Sex (female)	0.50	0.48	0.60	0.48

	Proportion	Proportion	Proportion	Proportion
Stroke[a,b]	0.18	0.05	0.20	0.08
PHT[b]	na	na	0.43	0.26

Characteristics of SCA patients in the primary study (CSSCD) and the validation study (Duke + BMC), and prevalence of complications. Superscripts a and b indicate the clinical variables that are significantly different between mild and severe patients in CSSCD (a) and Duke + BMC (b) (*P* value < 0.05). Note that LDH was measured on a different scale in the Duke and BMC patients compared to the CSSCD patients. However, in all studies, higher values of LDH are associated with increased severity.

TABLE II

SNPs that Reach Genome Wide Significance in the Discovery Set and Have Consistent Associations in the Validation Set

						CSSCD Set (1,265 SCA patients)				Validation Set (163 SCA patients)					
SNP	**Band**	**CEU**	**YRI**	**CSSCD**	**Allele**	**BF**	**OR**	***P* value**	***P*(*A*)**	**BF**	**OR**	***P* value**	***P*(*A*)**	**Gene**	**SSEA**
rs2745061	20q13.12	0.84	1.00	0.84	GvA	1,568,029	0.44	1.49E-09	0.70/0.84	0.56	0.67	0.157831	0.69/0.77		
rs6461625	7p15.3	0.53	0.65	0.64	GG v AA/AG	2,887	2.47	2.70E-06	0.26/0.13	0.66	1.78	0.300779	0.16/0.10		
rs13061797	3q26.1	0.86	0.82	0.86	GG v AA/AG	1,827	6.35	2.02E-06	0.07/0.01	0.54	1.40	0.923118	0.06/0.04		
rs2125053	10q23.2	0.91	0.58	0.56	AG/GG v AA	1,184	0.45	9.77E-06	0.68/0.83	0.29	0.82	0.746608	0.74/0.78	*BMPR1A*	0.18
rs4421324	8q22.1	0.91	0.83	0.81	GG v AA/AG	1,131	3.65	4.34E-06	0.15/0.04	0.53	1.39	0.930425	0.06/0.04		
rs12439075	15q26.1	0.63	0.83	0.81	GG v AA/AG	1,082	2.25	6.70E-05	0.81/0.66	0.42	1.47	0.376507	0.74/0.66	*AGC1*	0.04
rs680545	2p13.1	0.55	0.64	0.52	GG v AA/AG	2,354	2.13	5.92E-06	0.42/0.27	0.23	1.13	0.838519	0.32/0.29		
rs12609878	19q13.2	0.88	0.82	0.78	AC v AA CC v AA	1,625	2.22 0.95	1.60E-05	0.52/0.33 0.04/0.05	0.23	1.35 0.66	0.431789	0.38/0.30 0.07/0.11	*SPRED3*	0.05

List of SNPs that met genome-wide significance in the discovery set (CSSCD) using either an allelic association model, or a recessive, dominant, or general association model, and have consistent associations in the validation set but failed to reach statistical significance (BF < 1) with the current sample size. Column 1: SNP identifier from dbSNP. Column 2: cytogenetic band. Column 3–5: major allele frequencies in the CEPH and Yoruban from the HapMap, and the mild cases in the CSSCD set. Column 6: alleles A v B. Column 7: Bayes factor (BF) that gives the posterior odds for the reported model of association versus the null model of no association assuming prior odds = 1. Column 8: odds ratio (OR) for the alleles in columns 6. Column 9: *P* value from χ^2 test (2 *df* for genotype association and 1 *df* otherwise). Column 10: allele frequencies in severe/mild patients. Columns 11–14: Bayes factor, odds ratio, P value from χ^2 test and allele frequencies in the replication set. Column 15: gene name when the SNP is in a known gene, and SSEA probability score of the whole gene.

TABLE III

List of SNPs that Met Genome Wide Significance in the Discovery Set (CSSCD) and Are Replicated in the Validation Set

						CSSCD set (1,265 SCA patients)				Validation set (163 SCA patients)				
SNP	**Band**	**CEU**	**YRI**	**CSSCD**	**Allele**	**BF**	**OR**	***P* value**	***P*(*A*)**	**BF**	**OR**	***P* value**	***P*(*A*)**	**Gene**
rs12124726	1p36.13	0.54	0.79	0.76	CC v AA/AC	63,769	3.59	2.55E-08	0.18/0.06	1.03	2.19	0.240007	0.88/0.94	***OTUD3***
rs3004119	20p11.21	0.54	0.95	0.89	G v A	3,712	0.49	1.61E-06	0.80/0.89	3.01	0.48	0.03013	0.81/0.90	
rs7124828	11q21	0.8083	0.8879	0.73	AC/CC v AA	31,056	0.35	9.15E-07	0.39/0.65	161	0.09	0.002495	0.36/0.86	
rs274646	5p15.31	0.6864	0.6583	0.67	AG/GG v AA	11,844	0.36	1.14E-05	0.15/0.33	15.2	0.28	0.013515	0.08/0.24	
rs3843754	19q13.2	0.74	0.86	0.75	AG v AA GG v AA	13,52.5	1.51 0.69	1.84E-05	0.55/0.37 0.38/0.56	2.61	4.67 3.23	0.053015	0.45/0.33 0.51/0.53	***KCNK6*** ***YIF1B***

The columns legend is as in Table II.

TABLE IV

Genes Enriched of Significant SNPs

	Analysis in the CSSCD set				Validation in the Duke + BMC sets		
Official gene name	**Chromosome**	**Total SNP**	**Significant SNPs**	**SSEA Score**	**Tot SNP**	**Significant SNPs**	**SSEA Score**
TNKS	8	35	14	3.01E-10	28	3	0.181558885
CLASP1	4	31	12	8.70E-09	28	7	0.073648937
AGBL3	7	12	8	1.19E-08	10	4	0.043890234
TBC1D23	3	10	7	6.27E-08	4	2	0.102618951
MAP3K13	3	15	8	1.33E-07	14	4	0.106807925
C8A	1	16	8	2.53E-07	10	1	0.340847979
MED13L	12	22	9	3.68E-07	8	1	0.381633712
UCHL3	13	29	10	5.32E-07	29	2	0.103907526
QRFPR	4	24	9	8.77E-07	13	1	0.267613843
C8orf33	8	6	5	1.49E-06	6	5	0.000449556
DLG1	3	21	8	3.13E-06	16	6	0.020451989

List of 11 genes that were found enriched of statistical significant SNPs using SSEA, and replicate the associations in the validation set. Column 1: official gene name. Column 2–4 chromosome location, number of SNPs in each gene in the illumina 6.10 array and number of SNPs with Bayes Factor >3. Column 5: SSEA score. Columns 6–8: the total number of SNPs available in the same gene in the replication set, the total number of SNPs with Bayes factor >3, and the SSEA score. Note that the total number of SNPs per gene is different between the discovery and the replication sets because of the smaller sample size of the replication set and the filter of SNPs with unobserved genotypes.

TABLE V

Functional Annotation

System	Gene category	EASE score	Official gene symbol
GO molecular function	Protein binding	9.98E-04	*AKAP11; ANXA11; DLG1; FBXL2; IL1R2; MYO1B; NCOA1; NCOA3; OSTF1; RNF19; RTN4; SNTG1; TNKS; TP53BP1; UBP1; WASPIP; YAP1; ZNF193*
Interpro	E-class P450 Group IV	1.51E-03	*CYP2C18; CYP2C9; CYP4F8*
GO molecular function	Unspecific monooxygenase activity	9.57E-03	*CYP2C18; CYP2C9; CYP4F8*
GO biological process	Neuropeptide signaling pathway	1.12E-02	*CENTD1; RASSF6; RIN3; TAC1*
GO molecular function	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	1.20E-02	*BBOX1; CYP2C18; CYP2C9; CYP4F8*
Interpro	Cytochrome P450 enzyme	1.76E-02	*CYP2C18; CYP2C9; CYP4F8*
KEGG pathway	Gamma-Hexachlorocyclohexane degradation *Homo sapiens*	1.81E-02	*CYP2C18; CYP2C9; CYP4F8*
Chromosome	*Homo sapiens* 10q	2.10E-02	*ANXA11; CYP2C18; CYP2C9; PLAC9; PRKG1; SGPL1*
Subcellular localization	Cytoplasmic	4.09E-02	*DLG1; FBXL2; FPGS; RPL7; SNTG1; TP53BP1*
KEGG pathway	Biodegradation of Xenobiotics *Homo sapiens*	4.72E-02	*CYP2C18; CYP2C9; CYP4F8*
KEGG pathway	Fatty acid metabolism *Homo sapiens*	4.87E-02	*CYP2C18; CYP2C9; CYP4F8*

Results of the functional annotation analysis for the genes that had a SSEA score >0.001. The EASE score was computed using the EASE program for gene enrichment analysis.