Published in final edited form as:

Blood Cells Mol Dis. 2011 June 15; 47(1): 41–45. doi:10.1016/j.bcmd.2011.04.002.

Ancestry of African Americans with Sickle Cell Disease

Nadia Solovieff^a, Stephen W. Hartley^a, Clinton T. Baldwin^b, Elizabeth S. Klings^c, Mark T. Gladwin^d, James G. Taylor VI^e, Gregory J. Kato^e, Lindsay A. Farrer^{a,c}, Martin H. Steinberg^c, and Paola Sebastiani^a

^aDepartment of Biostatistics, Boston University School of Public Health, Boston, MA, USA

^bCenter for Human Genetics, Boston University School of Medicine, Boston, MA, USA

^cDepartment of Medicine, Boston University School of Medicine, Boston, MA, USA

^dDivision of Pulmonary, Allergy and Critical Care Medicine University of Pittsburgh Medical Center Pittsburgh, PA, USA

^eVascular Medicine Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA

Abstract

The inheritance of genetic disease depends on ancestry that must be considered when interpreting genetic association studies and can provide insights when comparing traits in a population. We compared the genetic profiles of African Americans with sickle cell disease to those of Black Africans and Caucasian populations of European descent and found that they are less genetically admixed than other African Americans and have an ancestry similar to Yorubans, Mandenkas and Bantu.

Keywords

sickle cell disease; genetic ancestry; admixture; genetic association

Introduction

Sickle cell disease (SCD) is caused by a point mutation in the β -hemoglobin gene (*HBB* glu6val) leading to the synthesis of sickle hemoglobin (HbS), which polymerizes when deoxygenated, distorting and injuring the red blood cell. Patients with sickle cell anemia are homozygous for this mutation and suffer from acute vasoocclusive events, hemolytic anemia, organ damage and failure, and an average lifespan reduction in developed countries of more than three decades [1]. The sickle hemoglobin mutation reached polymorphic

© 2011 Elsevier Inc. All rights reserved.

Corresponding Author: Nadia Solovieff, Boston University School of Public Health, Department of Biostatistics, 801 Massachusetts Avenue, Crosstown Building, 3rd Floor, Boston, MA, 02118 USA, ntimofoe@bu.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Authorship Contributions NS, SWH, developed the database worked on bioinformatics, performed all analyses and revised the manuscript. CTB performed SNP genotyping. LAF provided phenotype and genotype data on the MIRAGE Study subjects. ESK, MG, GK and JT provided patient information and edited the manuscript. PS, LAF, CTB, and MHS conceived the study, wrote the manuscript, analyzed data and revised the manuscript.

Disclosure of Conflicts of Interest: None

frequency in areas of Africa—other origins of the HbS gene were in the Middle East and Indian subcontinent—where malaria was prevalent, as carriers have a survival advantage and are more likely to survive to reproduce. In the United States, sickle cell anemia is most common among African Americans [1].

African Americans have from 20% to 30% genetic admixture with Europeans [2]. Studies using very limited genetic data have suggested that the HbS gene was more prevalent in African Americans with less Caucasian genetic admixture by examining the glucose-6-phosphate dehydrogenase (*G6PD*) genotypes in 48 subjects [3]. We used approximately 250,000 single nucleotide polymorphisms (SNPs) throughout the genome to examine the level of admixture and genetic ancestry of 1,810 African Americans with SCD and compared patients' genetic profiles to three groups of African Americans without SCD and several Black African and Caucasian European populations. African Americans with SCD have less Caucasian admixture than African Americans without the disease; their ancestry is most similar to the Yoruban, Mandenka and Bantu populations in Western Africa. These observations can be applied to the interpretation of genetic association studies, particularly when the trait of interest is related to the level of admixture.

Materials and Methods

Genome-wide SNP data were combined from 8 different datasets (Table 1). Patients with sickle cell disease were pooled from the Multicenter Study of Hydroxyurea (MSH) [4], NIH Pulmonary Hypertension Study [5] (NIH-PH), Boston University Pulmonary Hypertension Study (BU-PH) [6], and the Cooperative Study of Sickle Cell Disease (CSSCD) [7]. For comparison, we included African Americans without sickle cell disease by pooling subjects from the Multi-Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE) study [8], randomly selected subjects from the Illumina repository and the HapMap [9] panel of subjects with African ancestry in Southwest USA (ASW). As reference populations, we included 4 additional HapMap populations and 14 African and European populations from the Human Genome Diversity Project (HGDP) [10] including the HapMap Yoruba in Ibadan, Nigeria (YRI), Luhya in Webuye, Kenya (LWK), Maasai in Kinyawa, Kenya (MKK), the CEPH Utah residents with Northwestern European Ancestry (CEU), Bantu, Biaka, Mandenka, Mbuti pygmy, San, HGDP Yoruba, Adygei, Basque, French, Italian, Orcadian, Russian, Sardinian and Tuscan.

MSH, NIH-PH, BU-PH and MIRAGE samples were genotyped using the Illumina 370K array; CSSCD samples were genotyped on the Illumina 610K array; African Americans from the Illumina repository were genotyped on the Illumina 550K array; European and African populations from the Illumina repository were genotyped on the Illumina 650K array; HapMap populations had genotypes for approximately 1.5 million SNPs available. Autosomal SNPs appearing in all arrays were used for the ancestry analysis yielding 253,880 SNPs with a call rate greater than 95% and minor allele frequency greater than 5%. All subjects had a call rate greater than 93% and all individuals with an identity by descent estimate greater than 0.30 indicating first degree relatives were excluded from the analysis.

To examine the ancestry of these subjects, we performed a principal component analysis [11] (PCA) using the genome-wide set of SNPs, which passed quality controls for all subjects, and then applied a clustering algorithm to group individuals with similar ancestry [12]. PCA is the standard approach to detecting underlying population substructure using genome-wide SNP data and reduces the dimension of the data by summarizing the largest amount of variability in the data to the top axes of variation called principal components (PCs). The top PCs often correspond to ancestry or geographic origin. To understand the joint pattern formed by the PCs, we applied our clustering algorithm to the top PCs to form

clusters grouping individuals with similar ancestry together and separating individuals with differing ancestry. We also estimated the genetic distance between the clusters using the F_{st} statistic to further understand the relationship among the populations. F_{st} is defined as the proportion of genetic diversity due to allele frequency differences among populations and can be interpreted as the distance between populations [2].

The clustering algorithm is described in detail in [12]. Briefly, it first identifies the top ancestry informative PCs by inspecting a heatmap and a scree plot of the top 20 PCs for patterns (Figure 1). Next, subjects are clustered with respect to the top PCs using k-means clustering for a range of cluster sizes (2–30). The algorithm provides a scoring index for each cluster size to automatically identify the optimal number of clusters. The scoring index averages measures of the accuracy of the subjects' cluster assignments, the stability of k-means clustering, and the ability of k-means to maximize the distance between clusters. The PCA and F_{st} estimates were computed using the software EIGENSOFT [11]. We used Fisher's exact test to determine whether a higher percentage of African Americans without SCD compared to African Americans with SCD was more similar to Caucasians.

Results

The top 13 PCs were most informative (Figure 1) and the cluster algorithm created 12 distinct clusters of varying ancestry (Figure 2). Caucasian subjects (Clusters 1, 7 and 12, Table 2 and Figure 2) are clearly separated from African subjects and the first two principal components (PCs) capture the admixture among the African Americans (Figure 3). The Biaka, San, Mbuti pygmy and the majority of the MKK (Maasai in Kinyawa, Kenya) fall into separate clusters with little similarity with African Americans (Table 2, Figure 2). Cluster 8 includes the largest concentration of African Americans with and without SCD, 55% and 43% respectively, all of the Yoruban, Mandenka and part of the Bantu population (Table 2, cluster 8). The greater similarity of African Americans with these Western African groups is consistent with the historical record of the North American slave trade [13]. Clusters 2 and 5 contain most of the remaining African Americans with and without SCD, but none of the other African or Caucasian population (Table 2). These 2 groups, nevertheless, are genetically very similar to cluster 8 as measured by the F_{st} statistic ($F_{st} = 0.0001$) (Table 3). The remaining African Americans cluster into clusters 1, 7, and 12 all of which contain Caucasian populations indicating that these African Americans have a high level of Caucasian admixture. Cluster 1 shows admixture with Mediterranean populations (Basque, Italian, Tuscan and Sardinian), cluster 7 with Adygei subjects (southeastern Europe) and cluster 12 with northern European populations (French, Orcadian, Russian, HapMap CEU). A significantly higher percentage ($p < 0.0001$) of African Americans without SCD fall into these clusters (1,7,12) than African Americans with SCD, indicating that African Americans with SCD have less Caucasian admixture (Table 2). For this comparison we collapsed clusters 2–6 and 8–11, which do not contain any Caucasians from HGP, into one cluster.

Discussion

To our knowledge, this is the first analysis of population substructure in African Americans with sickle cell disease using genome-wide data. African Americans with SCD are less admixed than the general population of African Americans and their genetic substructure is most similar to Yorubans, Mandenkas and Bantus. We speculate that African Americans with SCD are less admixed because they must inherit two copies of the sickle mutation that is more common among African populations. Subjects with higher levels of Caucasian admixture are less likely to carry the sickle mutation and thus less likely to pass it to their offspring. An important consequence of our analysis is that Yorubans would be the most

appropriate reference panel of the HapMap population when imputing unknown or missing SNPs in genome-wide association studies in SCD.

Understanding genetic architecture can be vital for interpreting genetic association studies and provide new insights into phenotypic differences between populations. For example, in our studies of the modulation of fetal hemoglobin (HbF) in sickle cell anemia, we found that in the Southwestern Province of Saudi Arabia, SCD patients have *HBB*-gene cluster haplotypes of African origin with a distribution very similar to that of African Americans, yet have HbF levels almost twice as high [14]. When we examined the genetic population structure of these Saudi patients, it resembled Arab populations despite the presence of an "African" HbS gene [13]. In these two populations the commonality of *HBB* haplotypes coupled with the genetic distance between these populations suggested that the evolution of unique genetic modifiers or unknown environmental influences might account for the higher HbF in Saudi patients.

Acknowledgments

Supported by NIH grants R01 HL 87681 and R01 HL 068970 (MHS), R01 AG09029 and R01 AG025259 (LAF).

References

1. Steinberg, MH.; Forget, BG.; Higgs, DR.; Weatherall, DJ. Disorders of Hemoglobin; Genetics, Pathophysiology, and Clinical Management. Cambridge: Cambridge University Press; 2009.
2. Cavalli-Sforza, LL.; Menozzi, P.; Piazza, A. The history and geography of human genes. Princeton, N.J: Princeton University Press; 1994.
3. Beutler E, Johnson C, Powars D, West C. Prevalence of glucose-6-phosphate dehydrogenase deficiency in sickle-cell disease. The New England journal of medicine. 1974; 290:826–828. [PubMed: 4817837]
4. Steinberg MH, Lu ZH, Barton FB, Terrin ML, Charache S, Dover GJ. Fetal hemoglobin in sickle cell anemia: determinants of response to hydroxyurea. Multicenter Study of Hydroxyurea. Blood. 1997; 89:1078–1088. [PubMed: 9028341]
5. Gladwin MT, Sachdev V, Jison ML, Shizukuda Y, Plehn JF, Minter K, Brown B, Coles WA, Nichols JS, Ernst I, Hunter LA, Blackwelder WC, Schechter AN, Rodgers GP, Castro O, Ognibene FP. Pulmonary hypertension as a risk factor for death in patients with sickle cell disease. The New England journal of medicine. 2004; 350:886–895. [PubMed: 14985486]
6. Klings ES, Anton Bland D, Rosenman D, Princeton S, Odhiambo A, Li G, Bernard SA, Steinberg MH, Farber HW. Pulmonary arterial hypertension and left-sided heart disease in sickle cell disease: clinical characteristics and association with soluble adhesion molecule expression. American Journal of Hematology. 2008; 83:547–553. [PubMed: 18383329]
7. Gaston M, Rosse WF. The cooperative study of sickle cell disease: review of study design and objectives. The American Journal of Pediatric Hematology/Oncology. 1982; 4:197–201. [PubMed: 7114401]
8. Green RC, Cupples LA, Go R, Benke KS, Edeki T, Griffith PA, Williams M, Hipps Y, Graff-Radford N, Bachman D, Farrer LA, Group MS. Risk of dementia among white and African American relatives of patients with Alzheimer disease. JAMA : the journal of the American Medical Association. 2002; 287:329–336. [PubMed: 11790212]
9. C. International HapMap. The International HapMap Project. Nature. 2003; 426:789–796. [PubMed: 14685227]
10. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. Worldwide human relationships inferred from genome-wide patterns of variation. Science (New York, N.Y.). 2008; 319:1100–1104.
11. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38:904–909. [PubMed: 16862161]

12. Solovieff N, Hartley SW, Baldwin CT, Perls TT, Steinberg MH, Sebastiani P. Clustering by genetic ancestry using genome-wide SNP data. *BMC genetics*. :108.
13. Nagel, RL.; Embury, SH.; Hebbel, RP.; Mohandas, N.; Steinberg, MH. *Origins and Dispersion of the Sickle Gene, Sickle Cell Disease: Basic Principles and Clinical Practice*. New York: Raven Press; 1994. p. 353-380.
14. Alsultan, A.; Solovieff, N.; Aleem, A.; AlGahtani, FH.; Al-Shehri, A.; Osman, ME.; Kurban, K.; Bahakim, H.; Al-Momen, AK.; Chui, DHK.; Steinberg, MH. Fetal Hemoglobin In Sickle Cell Anemia: Molecular Characterization of Saudi Patients From the Southwestern Province; ASH Annual Meeting; p. 1621Abstracts

Figure 1. Heatmap, Scree Plot and Scoring Index

The heatmap (A) and scree plot (B) are used to select the top informative PCs while the scoring index (C) is used to select the number of optimal ethnic clusters. A heatmap graphically displays the PCs in a colored matrix in which the columns represent PCs, the rows represent subjects and the intensity of the color represents the PC value (green: higher than average, red: lower average). The rows of the heatmap are reordered using hierarchical clustering placing individuals with similar PC values next to each other. The scree plot graphs the natural logarithm of the eigenvalue for each PC which represents the amount of variability explained by the respective PC. A strong pattern is observed in the top 13 PCs suggesting that these are important and should be used to cluster the data. A kink is observed in the scree plot suggesting that the top 5 PCs are informative however we use the top 13 PCs for clustering because a pattern is observed for PCs 6–13 in the heatmap. The scoring index maximizes at 14 clusters (0.88, 95% CI: (0.86, 0.91) but 12 clusters (0.87, 95% CI: (0.84, 0.89)) is optimal since the scoring index for 12 falls within the 95% CI for 14 clusters indicating that it is not significantly better than 14 clusters.

Figure 2. Pair-wise PC Plots of Top 8 PCs

The 4 panels plot the top 8 PCs and are colored by the 12 cluster assignments produced by the clustering algorithm. The legend details the colored correspondence between cluster assignment and reported ancestry.

Figure 3. Plots of Principal Components 1 versus 2 by Study and Ethnic Origin

The 3 panels plot PC1 versus PC2 for the Hapmap and Hapmap colored by populations (A), the African Americans with SCD (B), and the African Americans without SCD (C). All 3 panels are plotted on the same scale and are comparable. PC1 differentiates the Africans from the Caucasians and shows the degree of admixture in the African Americans with and without SCD groups. PC2 differentiates the various ethnicities of African descent. The close proximity of the SCD and African Americans to the Yorubans (YRI), Mandenka and Bantu in the plot suggests a strong similarity to these populations located in western Africa.

Table 1
Populations and Studies

The name of the study, ethnic group and sample size is reported.

Population/Dataset	N
Human Genome Diversity Project: African	
Bantu	20
Biaka	32
Mandenka	24
Mbuti pygmy	15
Mozabite	30
San	6
Yoruba	24
Cooperative Study of Sickle Cell Disease	1460
Multicenter Study of Hydroxyurea in Sickle Cell Disease	176
NIH Pulmonary Hypertension Study	147
Boston University Pulmonary Hypertension	27
HapMap	
CEPH Utah Residents with NW European Ancestry (CEU)	60
Yoruba in Ibadan, Nigeria (YRI)	58
Luhya in Webuye, Kenya (LWK)	83
Maasai in Kinyawa, Kenya (MKK)	143
African ancestry in Southwest USA (ASW)	42
Random Set of African Americans from Illumina repository	246
MIRAGE	46
Human Genome Diversity Project: European	
Adygei	17
Basque	24
French	29
Italian	13
Orcadian	16
Russian	25
Sardinian	28
Tuscan	8

Table 2

12 Cluster Allocation

The clustering algorithm separated the subjects into 12 clusters. The values are row percentages representing the percentage of subjects in each group allocated to the 12 clusters.

Ethnic Group	Cluster											
	1	2	3	4	5	6	7	8	9	10	11	12
African	0.1	22.6	0	0	21.6	0	0.6	54.8	0	0	0.2	0.2
With SCD												
American	1.5	21.9	0	0	22.2	0	1.2	42.5	0	0	0	10.8
Without SCD												
African	0	0	0	0	0	0	0	40	0	0	60	0
Bantu												
Biaka	0	0	0	0	0	96.9	0	0	0	0	0	3.1
Luhya	0	0	0	0	0	0	0	0	0	0	100	0
Mandenka	0	0	0	0	0	0	0	100	0	0	0	0
Mbuti pygmy	0	0	100	0	0	0	0	0	0	0	0	0
Maasai	0	0	0	95.1	0	0	0	0	0	0	4.9	0
Mozabite	0	0	0	0	0	0	0	10	90	0	0	0
San	0	0	0	0	0	0	0	0	0	100	0	0
YRI	0	0	0	0	0	0	0	100	0	0	0	0
European	0	0	0	0	0	0	100	0	0	0	0	0
Adygei												
Basque	100	0	0	0	0	0	0	0	0	0	0	0
CEU	0	0	0	0	0	0	0	0	0	0	0	100
French	0	0	0	0	0	0	0	0	0	0	0	100
Italian	92.3	0	0	0	0	0	0	0	0	0	0	7.7
Orcadian	0	0	0	0	0	0	0	0	0	0	0	100
Russian	0	0	0	0	0	0	0	0	0	0	0	100
Sardinian	100	0	0	0	0	0	0	0	0	0	0	0
Tuscan	100	0	0	0	0	0	0	0	0	0	0	0

Table 3

Fst estimates between the 12 cluster allocation of subjects

Fst is a measure of genetic distance between two populations with higher values indicating greater distance. The Fst values support the conjecture that African populations have greater genetic diversity than Caucasian populations as demonstrated by the large Fst estimates between various African clusters relative to the smaller estimates between Caucasian clusters. Clusters 2 and 5 contain large numbers of African Americans but do not groups with any of the African populations; however, they are genetically very close to cluster 8 with Fst estimates of 0.001, indicating a strong similarity to the Yorubans, Mandenka and Bantu individuals in cluster 8. Note that these distances are smaller than the distances observed between clusters 1, 7 and 12 which include Caucasians of European descent. Clusters 2 and 5 are closer to the Caucasian clusters (1,7,12) than cluster 8, suggesting that the African Americans in clusters 2 and 5 are more admixed than those in cluster 8.

Populations	Cluster	1	2	3	4	5	6	7	8	9	10	11	12
European, AA, SCD	1	0	0.104	0.209	0.097	0.105	0.181	0.014	0.118	0.029	0.223	0.138	0.005
AA, SCD	2	0.104	0	0.078	0.015	0.001	0.042	0.07	0.001	0.067	0.1	0.008	0.098
Mbuti pygmy	3	0.209	0.078	0	0.085	0.077	0.058	0.17	0.076	0.167	0.099	0.072	0.203
Maasai	4	0.097	0.015	0.085	0	0.015	0.055	0.066	0.017	0.061	0.106	0.016	0.094
AA, SCD	5	0.105	0.001	0.077	0.015	0	0.042	0.07	0.001	0.068	0.1	0.008	0.099
Biaka	6	0.181	0.042	0.058	0.055	0.042	0	0.141	0.039	0.137	0.083	0.037	0.175
Adygei, AA, SCD	7	0.014	0.07	0.17	0.066	0.07	0.141	0	0.082	0.025	0.185	0.1	0.01
Bantu, Mandenka, Mozabite, YRI, AA, SCD	8	0.118	0.001	0.076	0.017	0.001	0.039	0.082	0	0.078	0.099	0.006	0.112
Mozabite	9	0.029	0.067	0.167	0.061	0.068	0.137	0.025	0.078	0	0.183	0.093	0.031
San	10	0.223	0.1	0.099	0.106	0.1	0.083	0.185	0.099	0.183	0	0.097	0.217
Bantu, Luhya, Maasai, SCD	11	0.138	0.008	0.072	0.016	0.008	0.037	0.1	0.006	0.093	0.097	0	0.133
European, AA, SCD	12	0.005	0.098	0.203	0.094	0.099	0.175	0.01	0.112	0.031	0.217	0.133	0