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], Paola Sebastiani [a]

[a] Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[b] Center for Human Genetics, Boston University School of Medicine, Boston, MA, USA
[c] Department of Medicine, Boston University School of Medicine, Boston, MA, USA
[d] Division of Pulmonary, Allergy and Critical Care Medicine University of Pittsburgh Medical Center Pittsburgh, PA, USA
[e] Vascular Medicine Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA

ARTICLE INFO

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.

© 2011 Elsevier Inc. All rights reserved.

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 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 1810 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,

* Corresponding author at: Boston University School of Public Health, Department of Biostatistics, 801 Massachusetts Avenue, Crosstown Building, 3rd Floor, Boston, MA, 02118 USA.

E-mail address: ntimofee@bu.edu (N. Solovieff).

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

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 Fst statistic to further understand the relationship among the populations. Fst 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 Ref. [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 (Fig. 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 Fst 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 (Fig. 1) and the cluster algorithm created 12 distinct clusters of varying ancestry (Fig. 2). Caucasian subjects (Clusters 1, 7 and 12, Table 2 and Fig. 2) are clearly separated from African subjects and the first two principal components (PCs) capture the admixture among the African Americans (Fig. 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, Fig. 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 Fst statistic (Fst = 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, and 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 HGDP, 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