ARTICLE

Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sickle Allele during the Holocene Wet Phase

Daniel Shriner[1] and Charles N. Rotimi[1,*]

Five classical designations of sickle haplotypes are made on the basis of the presence or absence of restriction sites and are named after the ethno-linguistic groups or geographic regions from which the individuals with sickle cell anemia originated. Each haplotype is thought to represent an independent occurrence of the sickle mutation rs334 (c.20A>T [p.Glu7Val] in *HBB*). We investigated the origins of the sickle mutation by using whole-genome-sequence data. We identified 156 carriers from the 1000 Genomes Project, the African Genome Variation Project, and Qatar. We classified haplotypes by using 27 polymorphisms in linkage disequilibrium with rs334. Network analysis revealed a common haplotype that differed from the ancestral haplotype only by the derived sickle mutation at rs334 and correlated collectively with the Central African Republic (CAR), Cameroon, and Arabian/Indian haplotypes. Other haplotypes were derived from this haplotype and fell into two clusters, one composed of Senegal haplotypes and the other composed of Benin and Senegal haplotypes. The near-exclusive presence of the original sickle haplotype in the CAR, Kenya, Uganda, and South Africa is consistent with this haplotype predating the Bantu expansions. Modeling of balancing selection indicated that the heterozygote advantage was 15.2%, an equilibrium frequency of 12.0% was reached after 87 generations, and the selective environment predated the mutation. The posterior distribution of the ancestral recombination graph yielded a sickle mutation age of 259 generations, corresponding to 7,300 years ago during the Holocene Wet Phase. These results clarify the origin of the sickle allele and improve and simplify the classification of sickle haplotypes.

Introduction

Several hereditary variants in the hemoglobin genes afford protection against malaria (MIM: 611162). Many such variants are thought to have evolved in the last 10,000 years.[1,2] In the beta globin gene *HBB* (MIM: 141900), the sickle (MIM: 603903) allele β^S is under balancing selection because of heterozygote advantage, which results from protection against malaria caused by *Plasmodium falciparum* and recessive lethality. The chromosomal background of the β^S allele has been classified on the basis of the presence or absence of a set of seven canonical restriction sites, 5′ ε HincII → Gγ1 HindIII → Aγ1 HindIII → ψβ HincII → 3′ ψβ HincII → β AvaII → 3′ β BamHI, yielding five haplotypes, Arabian/Indian, Benin, Cameroon, CAR, and Senegal, named after ethno-linguistic groups or geographic regions, as well as a sixth category for "atypical" haplotypes.[3–9] These designations do not necessarily indicate where the haplotypes originated.

Whether the β^S allele has a recent or old origin has been debated since the development of restriction fragment length polymorphism (RFLP) data.[10,11] According to the multicentric model, the origin of the β^S allele is recent—within the last few thousand years—and each haplotype represents an independent occurrence of the same exact mutation in the corresponding geographic region.[4,12–14] In contrast, according to the unicentric model, the origin of the β^S allele is anywhere from tens to hundreds of thousands of years old, and the mutation occurred once.[15–18] Suggested places of origin include equatorial Africa[19] and the Middle East.[20,21] A 1.2 kb recombination hotspot exists 1 kb upstream of *HBB*.[22] Consequently, recombination and gene conversion, rather than *de novo* mutation, have generated several haplotypes.[3,23,24]

We investigated the origins of the sickle allele by using whole-genome-sequence data from a total of 2,932 individuals from the 1000 Genomes Project, the African Genome Variation Project, and Qatar. We identified a total of 156 sickle carriers. We classified haplotypes by using phased sequence data comprising diallelic sites (both single-nucleotide polymorphisms and short insertions or deletions). We then used a combination of forward time simulation, phylogenetic network analysis, and coalescent analysis to infer a single origin of the sickle allele approximately 7,300 years ago, during the Holocene Wet Phase or Green Sahara.

Material and Methods

Ethics Statement

This project was excluded from institutional-review-board review by the Office of Human Subjects Research Protections, National Institutes of Health (OHSRP ID# 17-NHGRI-00282).

Sequence Data

We retrieved whole-genome-sequence data from 2,504 individuals in the 1000 Genomes Project,[25] 320 individuals in the African Genome Variation Project,[26] and 108 individuals from Qatar.[27] As previously detailed by the 1000 Genomes Project, a three-stage approach was used to establish phased haplotypes: (1) given

[1] Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, MD 20892, USA
*Correspondence: rotimic@mail.nih.gov

https://doi.org/10.1016/j.ajhg.2018.02.003.

array-based genotype data and family information, SHAPEIT2 was used to estimate a scaffold of long-range phased haplotypes, (2) a combination of Beagle and SHAPEIT2 was used to jointly analyze the long-range phased haplotypes and di-allelic variants, and (3) MVNcall was used to place multi-allelic and structural variants onto the haplotype scaffold.[25] The average switch error rate was 0.56%, and the average distance between phasing errors was 1,062 kb.[25] For comparison, the length of the β-globin locus, including the locus control region, is 67 kb. All data were accessed with VCFtools version 0.1.14.[28] We assessed linkage disequilibrium in the 1000 Genomes Project data by using the --hap-r2 function for phased haplotypes.

Measuring Multi-locus Association

We calculated Cramér's V for the association between two nominal variables. In our analyses, the number of levels equaled the number of haplotypes. V^2 is the square of the mean canonical correlation and is equivalent to Pearson's r^2 if at least one variable has only two levels. We implemented a bias-corrected version of Cramér's V given by $\tilde{V} = \sqrt{\tilde{\varphi}^2/\tilde{m}}$, $\tilde{\varphi}^2 = \max(0, \varphi^2 - ((r-1)(c-1)/n-1))$, $\varphi^2 = \chi^2/n$, $\tilde{m} = \min(\tilde{r}-1, \tilde{c}-1)$, $\tilde{r} = r - ((r-1)^2/n-1)$, $\tilde{c} = c - ((c-1)^2/n-1)$, where n is the sample size and χ^2 is the chi-square statistic without a continuity correction for a contingency table with r rows and c columns.[29]

African Ancestry

We used YFitter to call Y chromosome haplogroups[30] and HaploFind to call mitochondrial DNA haplogroups.[31] We used projection analysis in ADMIXTURE version 1.3[32] with a reference panel of 21 global ancestries[33] to analyze autosomal ancestry. To determine standard errors for the proportions of ancestral components for each individual, we reran ADMIXTURE with the addition of 200 bootstrap replicates. Accounting for both within and between individual variances, we calculated the proportions for average ancestry by using inverse variance weights. We then calculated 95% confidence intervals for each ancestry and individual, zeroed out any average proportions for which the 95% confidence intervals included 0, and renormalized the remaining averages to sum to 1.

Balancing Selection

Let the genotype frequencies of the sickle homozygote, heterozygote, and wild-type homozygote be p^2, $2pq$, and q^2, respectively. Let the corresponding relative fitnesses be 0, $1+s$, and 1, respectively.[2] Then, at equilibrium, $s = (p/1-2p)$. For each of the five continental African samples in the 1000 Genomes Project Phase 3 release version 5a, we estimated the effective population size N_e on the basis of the heterozygosities of all single-nucleotide polymorphisms (i.e., diallelic, triallelic, and quadrallelic) by assuming a mutation rate of 0.97×10^{-8} mutations per site per generation.[34] We then took the harmonic mean of the five N_e estimates. Assuming one initial copy of the mutant allele, we simulated 1,000 generations under a combination of random genetic drift and balancing selection. We repeated this process 1,000 times.

Phylogenetic Network Analysis

Phylogenetic trees are based on models of evolution that assume a purely bifurcating process. This assumption is violated by several evolutionary processes, including recombination, gene conversion, and recurrent mutation. A single phylogenetic tree is unable to represent incompatible signals across sites. In contrast, a phylogenetic network is a more general form of graph that relaxes the assumption of bifurcation. Split decomposition is an inferential method that computes a set of incompatible splits from a given distance matrix. Trivial splits separate a set of taxa into two sets, one set containing a single taxon and the other set containing all other taxa, thereby defining terminal branches. Non-trivial splits separate a set of taxa into two sets, both containing at least two taxa, thereby defining internal branches. A split network depicts incompatible splits as parallel branches. We used the Neighbor-Net method implemented in SplitsTree version 4.13.1 to perform split decomposition analysis of haplotypes.[35]

Inferring the Ancestral Recombination Graph

We used ARGweaver to infer the ancestral recombination graph.[36] ARGweaver is based on the standard coalescent model and is sensitive to balancing selection, such that regions under balancing selection have older times to the most recent common ancestor than comparable neutral regions. To account for uncertainty in both the mutation and recombination rates, we used a grid approach. We tested mutation rates of 0.72×10^{-8}, 0.97×10^{-8}, and 1.44×10^{-8} mutations per generation per site.[34] We tested recombination rates of 5×10^{-9}, 1×10^{-8}, 1.5×10^{-8} (the default value for human data), 2×10^{-8}, and 2×10^{-7} recombinations per generation per site.[34] We ran the sampler for 3,000 iterations and saved the ancestral recombination graph from every tenth iteration. We used the functions heidel.diag and geweke.diag in the coda library of R, version 3.2.3, to assess convergence diagnostics on the basis of the posterior distribution of the number of recombination events.[37] To convert generations into years, we assumed a generation interval of 28 years.[38,39]

Results

Molecular Mapping of Restriction Sites

We mapped 15 restriction sites, including the seven canonical sites, to the reference human genome sequence. We identified 12 known markers that predict the presence or absence of ten of these sites. Of the canonical sites, we predicted 5′ ε HincII with rs3834466, Gγ1 HindIII with rs2070972, Aγ1 HindIII with rs28440105, and 3′ ψβ HincII with rs968857 (Table 1); similar results were described in a previously reported analysis of rs3834466, rs28440105, rs10128556, and rs968857.[40] Pairwise correlation (measured via r^2) between these variants and rs334 was weak to nonexistent (Table 1). On the basis of these four RFLP-predicting markers, we identified ten unique haplotypes. $\tilde{V}^2$ between the set of ten haplotypes and rs334 was 0.079 (Figure 1).

Distributions of β^S and the Classical Haplotypes

In the 1000 Genomes Project data, we identified 137 sickle carriers and 0 sickle homozygotes; we predicted the classical haplotypes for all 137 carriers (Table 2). The Benin haplotype was the predominant haplotype in the samples of Esan and Yoruba from Nigeria, the CAR haplotype was the predominant haplotype in the sample of Luhya from Kenya, and the Senegal haplotype was the predominant haplotype in the samples of Mende from Sierra Leone and Mandinka from the Gambia. There was limited

Table 1. Molecular Characterization of the Classical Sickle Haplotypes

	Site			
	3′ ψβ HincII	Aγ1 HindIII	Gγ1 HindIII	5′ ε HincII
Sequence[a]	GTTGAC	AAGCTT	AAGCTT	GTTGAC
Range	5,260,457– 5,260,462	5,269,799– 5,269,804	5,274,717– 5,274,722	5,291,563– 5,291,567
rsID	rs968857	rs28440105	rs2070972	rs3834466
Position (hg19)	5,260,458	5,269,799	5,274,717	5,291,563– 5,291,564
Senegal	+	−	+	−
Benin	+	−	−	−
CAR	−	−	+	−
Cameroon	+	+	+	−
Arabian/Indian	+	−	+	+
r^2[b]	0.000	0.016	0.003	0.020
D′[b]	−0.104	0.930	0.094	−0.853
Ancestral	T	C	C	G
Status	+	−	−	−
Derived	C	A	A	GT
Status	−	+	+	+

[a] Underlining indicates the polymorphic position.
[b] Pairwise linkage disequilibrium values are shown with respect to rs334.

representation of the Arabian/Indian (one) and Cameroon (two) haplotypes. The average sickle allele frequency was 12.0% and did not statistically differ among the five continental African samples ($\chi^2_4 = 1.48$, $p = 0.830$; Table S1). The distribution of matrilines comprised 1 A2, 2 B2, 1 J2, 8 L0, 24 L1, 43 L2, 49 L3, 3 L4, 2 L5, 1 T2, and 3 U6 haplogroups. The distribution of patrilines comprised 2 A1a, 5 E1a, 54 E1b1a, 1 E1b1b, 2 E2b, 1 G2a, 1 I2a, and 2 R1b haplogroups. Of the 54 E1b1a, 29 were E1b1a1a1f and 23 were E1b1a1a1g.

In the African Genome Variation Project data, we identified 14 sickle carriers in the Baganda and one sickle carrier in the Zulu. We predicted that all 15 of these individuals carried the CAR haplotype (Table 2). In the Qatar sample, we identified four sickle carriers, all with insufficient information to predict the haplotypes.

Haplotype Classification Based on Linkage Disequilibrium

We defined haplotypes centered on rs334 in the 504 continental Africans from the 1000 Genomes Project. We recorded pairwise linkage disequilibrium (LD) between rs334 and all phased, diallelic markers on chromosome 11 (Figure 2). The largest value of r^2 with rs334 was 0.407. There was only one marker, rs149481026, with $r^2 \geq 0.4$. This one marker was more strongly associated with rs334 ($\tilde{V}^2 = 0.406$) than the set of four RFLP-predicting markers was (Figure 1A).

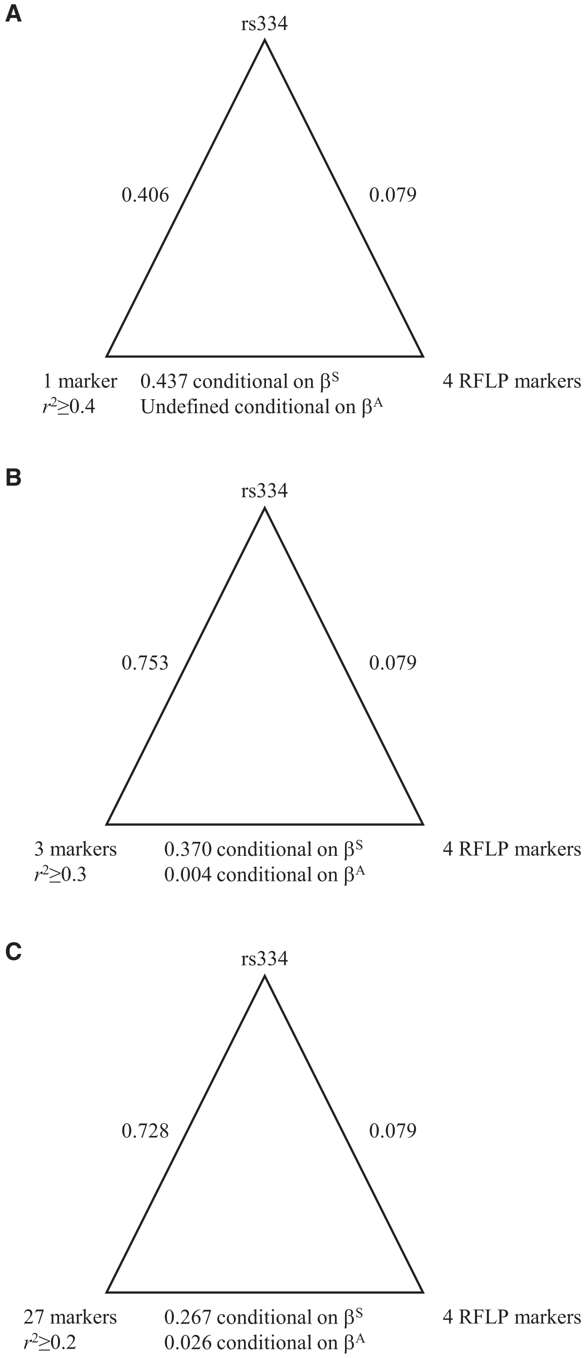
A
rs334
0.406
0.079
1 marker
$r^2 \geq 0.4$
0.437 conditional on β^S
Undefined conditional on β^A
4 RFLP markers
B
rs334
0.753
0.079
3 markers
$r^2 \geq 0.3$
0.370 conditional on β^S
0.004 conditional on β^A
4 RFLP markers
C
rs334
0.728
0.079
27 markers
$r^2 \geq 0.2$
0.267 conditional on β^S
0.026 conditional on β^A
4 RFLP markers

Figure 1. Pairwise Association Plots
For each triangle, vertices indicate markers, and edges are labeled with $\tilde{V}^2$ association values. The associations on the bottom edges are conditional on the presence of the β^S or β^A allele.
(A) Associations between rs334, the RFLP-predicting markers, and the one marker with pairwise $r^2 \geq 0.4$ with rs334.
(B) Associations between rs334, the RFLP-predicting markers, and the set of three markers with pairwise $r^2 \geq 0.3$ with rs334.
(C) Associations between rs334, the RFLP-predicting markers, and the set of 27 markers with pairwise $r^2 \geq 0.2$ with rs334.

To strengthen the association with rs334, we investigated lower levels of r^2. There were three markers with $r^2 \geq 0.3$: rs183055323, rs149481026, and rs73404549. These three markers ranged across 132.6 kb, including

Table 2. Distribution of Classical Sickle Haplotypes

Sample[a]	Arabian/Indian	Benin	Cameroon	CAR	Senegal	Atypical
ACB	0	4	0	2	3	0
ASW	0	1	1	0	0	0
CLM	0	0	0	1	0	1
ESN	0	18	1	0	5	0
GWD	0	2	0	0	24	0
LWK	1	0	0	19	0	0
MSL	0	3	0	0	17	1
PUR	0	0	0	1	2	0
YRI	0	19	0	0	10	1
Baganda	0	0	0	14	0	0
Zulu	0	0	0	1	0	0

[a] The population codes are as follows: ACB, African Caribbean in Barbados; ASW, People with African Ancestry in Southwest USA; CLM, Colombians in Medellín, Colombia; ESN, Esan in Nigeria; GWD, Gambian in Western Division, Mandinka; LWK, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; PUR, Puerto Ricans in Puerto Rico; and YRI, Yoruba in Ibadan, Nigeria.

the entire β-globin cluster. This interval was substantially smaller than the average distance between phasing errors. On the basis of these three markers, we identified five unique haplotypes (Table 3). One haplotype contained all occurrences of the Arabian/Indian, Cameroon, and CAR haplotypes. This haplotype contained the ancestral allele at all three markers. The Senegal haplotype was distributed across all five haplotypes, and the Benin haplotype was distributed across four of the five haplotypes. $\tilde{V}^2$ between these three markers and rs334 was 0.753 (Figure 1B).

To improve cross-classification with the Senegal and Benin haplotypes, we identified 27 markers with $r^2 \geq 0.2$. These markers extended across 725.3 kb, which is still less than the average distance between phasing errors. On the basis of these markers, we identified 59 unique haplotypes, of which 18 carried the sickle allele at rs334. A 19th sickle haplotype was observed once in the ACB sample, and a 20th sickle haplotype was observed once in the Baganda sample (Table 4). $\tilde{V}^2$ between these 27 markers and rs334 was 0.728 (Figure 1C). The most common haplotype carried the ancestral allele at all 28 sites and accounted for 68.5% of all haplotypes in the continental Africans. Globally, the ancestral haplotype had a frequency of 91.9%, was the most frequent haplotype in all 26 samples in the 1000 Genomes Project, and was the only haplotype in 15 of those samples, including all ten samples from East Asia and Europe. 13 of the sickle haplotypes in the Baganda, the one in the Zulu, and all four in the Qatari were identical to HAP1, the haplotype most commonly designated CAR (Table 4). Additionally, the four Qatari carriers had (1) >7.8% autosomal African ancestry, (2) an African mitochondrial haplogroup, or (3) an African Y chromosome haplogroup. The three most common haplotypes (HAP1, HAP16, and HAP20) correlated primarily with the CAR, Benin, and Senegal haplotypes, respectively (Table 4).

Given that our sets of 1, 3, and 27 markers more strongly associated with rs334 than did the set of RFLP-predicting markers, we assessed association between our sets of markers and the set of RFLP-predicting markers. Our sets of markers were moderately associated with the set of RFLP-predicting markers when conditioned on the presence of β^S and weakly associated when conditioned on the presence of β^A (Figure 1).

Bioinformatic Annotation

Different haplotypes might be associated with different clinical phenotypes or disease severity.[41] Using Ensembl and HaploReg version 4.1,[42] we annotated each of the 27 markers in linkage disequilibrium at $r^2 \geq 0.2$ with rs334 (Table S1). Nine sites were marked on histones as promoters or enhancers, possibly implicating differential expression of *HBB*, *HBD* (MIM: 142000), *HBE1* (MIM: 142100), or *HBG2* (MIM: 142250). Three of these nine sites, rs334, rs73402608, and rs112336602, were also annotated as bound by proteins, possibly implicating the same four genes. In addition, rs1039215 is correlated with gene expression, most strongly with *HBG2* (Table S2).

Balancing Selection

We modeled balancing selection assuming that the relative fitness of the β^A/β^A homozygote was 1, the relative fitness of the β^S/β^S homozygote was 0, and the relative fitness of the β^A/β^S heterozygote was $1 + s$. On the basis of the 504 continental Africans from the 1000 Genomes Project, we estimated that $s = 0.158$, which was in agreement with previous estimates.[14] Next, by assuming a single initial copy and an effective population size $N_e = 25{,}542$, we modeled random genetic drift plus balancing selection to

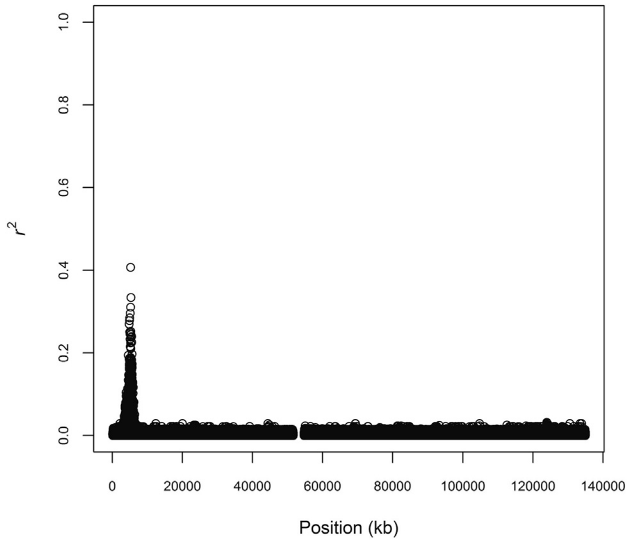

Figure 2. Linkage Disequilibrium with rs334
We calculated pairwise r^2 with rs334 across chromosome 11 among 504 continental Africans from the 1000 Genomes Project. We plotted r^2 for all 4,024,958 phased diallelic markers.

estimate how many generations it would take for an equilibrium frequency of 12.0% to be reached. We found that the mutant allele was lost 74.6% of the time and, conditional on reaching equilibrium, reached a frequency of 12.0% after an average of 87 (95% confidence interval [68, 124]) generations, or approximately 2,400 (95% confidence interval from 1,900 to 3,500) years. We stress that this value is not the age of the sickle mutation nor the age since the onset of balancing selection but rather the time to reach a frequency of 12.0%. To determine the fate of the mutant allele in the absence of heterozygote advantage, we repeated the simulation while assuming $s = 0$. We found that the mutant allele was lost after an average of 12 generations (95% confidence interval [1, 92]), with a median of two generations.

Phylogenetic Network Analysis

We used split decomposition analysis to infer the phylogeny of the 20 sickle haplotypes, whereby the phylogeny was rooted by the ancestral haplotype. (Figure 3). The network revealed that the sickle mutation occurred once in the background of the ancestral haplotype and gave rise to HAP1, which is associated predominantly with the CAR haplotype. One cluster containing HAP1 and HAP10 accounted for all occurrences of the Cameroon, CAR, and Arabian/Indian haplotypes (Table 4). Two additional clusters were derived from this haplotype and not from the ancestral haplotype. One cluster contained HAP6, HAP9, HAP19, and HAP20; all associated with the Senegal haplotype (Table 4). In this cluster, the modal haplotype, HAP20, differed from HAP1 by 15 derived mutations (Table 4 and Table S1). This cluster accounted for 70.5% of occurrences of the Senegal haplotype. The other cluster contained the remaining 14 haplotypes, accounting for all occurrences of the Benin haplotype and 29.5% of the occurrences of the Senegal haplotype. In this cluster, the modal haplotype, HAP16, differed from HAP1 by 13 derived mutations (Table 4; Table S1). Other than the sickle mutation, only one derived mutation, at rs10655224, was shared between HAP16 and HAP20. The remaining 14 derived mutations in HAP20 had an average frequency of 11.3% in the Gambian in Western Divisions in the Gambia (GWD) and Mende in Sierra Leone (MSL) samples and 0.1% in the Esan in Nigeria (ESN) and Yoruba in Ibadan, Nigeria (YRI) samples (Table S1). Conversely, the remaining 12 derived mutations in HAP16 had an average frequency of 13.4% in the ESN and YRI samples and 2.2% in the GWD and MSL samples (Table S1). These 26 derived alleles had an average frequency of 0.4% in the Luhya in Webuye, Kenya (LWK) sample (Table S1).

The Ancestral Recombination Graph

Using coalescent theory, we sampled the posterior distribution of the ancestral recombination graph with the 1,008 haplotypes, including 121 sickle haplotypes, from the 504 continental Africans from the 1000 Genomes Project. Over a grid of 15 pairs of mutation and recombination rates, a mutation rate of 0.97×10^{-8} mutations per generation per site and a recombination rate of 1.5×10^{-8} recombinations per generation per site yielded the best fit to the data. Given these two rates, the trace of the posterior number of recombination events visually indicated convergence to a stationary distribution (Figure S1). More formally, Geweke's diagnostic indicated that the sampled values came from a stationary distribution ($p = 0.063$). Heidelberger and Welch's diagnostic also indicated that the sampled values came from a stationary distribution ($p = 0.163$) and further indicated that there was no need to discard initial iterations. We estimated the age of the sickle mutation as 259 (95% credible interval [123, 395]) generations, or approximately 7,300 years (95% credible interval from 3,400 to 11,100 years).

Discussion

There are two models of the origins of the sickle allele. The multicentric model posits five independent occurrences of the same mutation within the last few thousand years. The unicentric model posits a single occurrence and an older age. We used whole-genome-sequence data to provide insight into this issue. Using haplotypic classification and phylogenetic network analysis, we found clear and consistent evidence for a single origin of the sickle mutation. After accounting for recombination, we estimated that the sickle mutation is 259 [123, 395] generations old.

Starting from first principles of population genetics, we defined haplotypes on the basis of phased sequence data. In contrast, the classical haplotypes were based on patterns of restriction sites seen in individuals with sickle cell anemia.[3,4] By molecularly mapping restriction sites to

Table 3. Distribution of Sickle Haplotypes under a Sequence-Based Classification Scheme Using Three Markers

Name	Haplotype[a]	Arabian/Indian	Benin	Cameroon	CAR	Senegal	Atypical
Ancestor	$00\underline{0}0$	NA	NA	NA	NA	NA	NA
HAPA	$00\underline{1}0$	1	4	2	38	1	1
HAPB	$00\underline{1}1$	0	2	0	0	1	0
HAPC	$01\underline{1}0$	0	1	0	0	1	1
HAPD	$01\underline{1}1$	0	40	0	0	15	0
HAPE	$10\underline{1}0$	0	0	0	0	43	1

[a] 0 indicates the reference allele, and 1 indicates the alternate allele according to the coding scheme in the 1000 Genomes Project VCF files. The sickle site rs334 is underlined.

the sequence data, we found that the restriction sites correlated poorly with rs334, such that none of the canonical sites were included in our sequence-based haplotypes. This discrepancy can be explained by the fact that the RFLP-defining markers circulate on both β^S- and β^A-carrying chromosomes, even though the restriction sites were originally used for defining types of β^S-carrying chromosomes. In contrast, our haplotypes correlate strongly with rs334 and simultaneously correlate moderately with the restriction sites conditional on the presence of the β^S allele, thereby capturing the classical patterns.

There are four lines of evidence regarding the age of the sickle mutation: historical data, the simulations of balancing selection, the patriline data, and the analysis of the ancestral recombination graph. Although historical records are sparse, the earliest recorded cases of fevers that could have been caused by malaria were ~5,000 years ago in China and were possibly due to *Plasmodium vivax*.[2,43,44] The earliest recorded case of illness that could have been malaria specifically caused by *Plasmodium falciparum* could have been ~4,000 years ago in Egypt and Sumer; however, *Plasmodium falciparum* could have been present in Africa several thousand years earlier.[2] The first recorded cases of sickle cell anemia or, more broadly, sickle cell disease were in Egypt during the predynastic period (~3200 BC[45]), in the Persian Gulf during the Hellenistic period (2,130 years before present[46]), and in Ghana in 1670 AD.[47] The existence of the sickle allele in predynastic Egypt constrains the lower bound of the age of the sickle mutation to be 5,200 years. On the basis of these limited data, it is historically plausible that the selective environment preceded the sickle mutation, consistent with our balancing-selection simulations, which showed that the sickle allele would have been lost almost immediately without a heterozygote advantage (under the assumption of recessive lethality).

The patriline data corroborate the age of the sickle mutation. Given a single origin, the sickle mutation most likely arose either in a population containing both E1b1a1a1f-L485 (also known as E1b1a1a1a1c-L485) and E1b1a1a1g-U175 (also known as E1b1a1a1a2a-U175) or in a population containing their common ancestor. Thus, the time of origin of the common ancestor of these two haplogroups is a plausible upper bound on the age of the sickle mutation. The common ancestor E1b1a1a1a-M4732 arose 10,500 (95% confidence interval between 9,200 and 12,000) years ago, consistent with the upper credible interval of 11,100 years ago that we estimated for the age of the sickle mutation.

The Bantu expansions started ~5,000 years ago and took ~2,000 years to cross from present-day Cameroon to the Great Lakes region.[48] The wide credible interval (3,400 to 11,100 years) of the age of the sickle mutation does not rule out the possibility that the mutation coincided with the onset of the Bantu expansions. However, our results support the hypothesis that the origin of the β^S allele predated the onset of the Bantu expansions. First, the β^S allele could have been at or near its equilibrium frequency when the Bantu expansions occurred ~5,000 years ago, according to our estimation that the sickle mutation originated ~7,300 years ago and that balancing selection ensued immediately and reached equilibrium in ~2,400 years in an environment where malaria preexisted. Second, the likelihood that the β^S allele is carried by migrating individuals increases with the frequency of that allele in the source population.

Our results provide some suggestions as to where the sickle mutation might have originated. Descendants of the Y chromosome haplogroup E1b1a-V38 migrated across the Sahara from east to west,[49] possibly around 19,000 years ago.[50] E1b1a1-M2 most likely did not originate in eastern or northeastern Africa, but where it originated in either western or central Africa is unclear.[49] Accordingly, the sickle mutation most likely did not occur in eastern or northeastern Africa. Our results indicate that the origin of the sickle mutation was in the middle of the Holocene Wet Phase, or Neolithic Subpluvial, which lasted from ~7,500–7,000 BC to ~3,500–3,000 BC. This time was the most recent of the Green Sahara periods, during which the Sahara experienced wet and rainy conditions.[51] Our results thus support the Green Sahara as a possible place of origin of the sickle mutation. An alternative hypothesis is that the sickle allele arose in west-central Africa,[13,52] possibly in the northwestern portion of the equatorial rainforest.

Table 4. Distribution of Sickle Haplotypes under a Sequence-Based Classification Scheme Using 27 Markers

Name	Haplotype[a]	Arabian/Indian	Benin	Cameroon	CAR	Senegal	Atypical
Ancestor	00000000000000000000$\underline{0}$0000010	NA	NA	NA	NA	NA	NA
HAP1	00000000000000000000$\underline{1}$0000010	1	0	2	37	0	1
HAP2	00000000000000000000$\underline{1}$0110010	0	4	0	0	0	0
HAP3	00000000000000000000$\underline{1}$0110101	0	0	0	0	1	0
HAP4	00000000000000000011$\underline{1}$0110101	0	2	0	0	0	0
HAP5	00000000000000001111$\underline{1}$0110101	0	5	0	0	0	0
HAP6	00000000000000011000$\underline{1}$1001010	0	0	0	0	3	0
HAP7	00000001000001001111$\underline{1}$0110101	0	1	0	0	0	0
HAP8	00000001000001001111$\underline{1}$0110110	0	0	0	0	1	0
HAP9	00001110111110111000$\underline{1}$1001010	0	0	0	0	1	0
HAP10	00010000000000000000$\underline{1}$0000010	0	0	0	1	0	0
HAP11	00010001000001001111$\underline{1}$0110101	0	2	0	0	1	0
HAP12	00110001000001001011$\underline{1}$0000010	0	0	0	0	1	0
HAP13	00110001000001001111$\underline{1}$0000010	0	1	0	0	0	1
HAP14	00110001000001001111$\underline{1}$0110010	0	0	0	0	1	0
HAP15	00110001000001001111$\underline{1}$0110100	0	2	0	0	1	0
HAP16	00110001000001001111$\underline{1}$0110101	0	23	0	0	7	0
HAP17	00110001000001001111$\underline{1}$0110110	0	6	0	0	5	0
HAP18	00110001000001001111$\underline{1}$0110111	0	1	0	0	0	0
HAP19	11001110111110111000$\underline{1}$0000010	0	0	0	0	3	1
HAP20	11001110111110111000$\underline{1}$1001010	0	0	0	0	36	0

[a] 0 indicates the reference allele, and 1 indicates the alternate allele according to the coding scheme in the 1000 Genomes Project VCF files. The sickle site rs334 is underlined.

Our results also provide some suggestions as to where the three clusters might have originated. Two splits occurred early in the original β^S-carrying population. The first split defined one cluster containing HAP1 and accounting for the Cameroon and CAR haplotypes. It is plausible that HAP1 was carried from an area in or around present-day Cameroon to the area that is presently the CAR, as well as to areas east and south, as part of the Bantu expansions. However, the Bantu expansions did not extend west and north. The second split subsequently separated the clusters containing HAP16 and HAP20, the modal haplotypes accounting for the Benin and Senegal haplotypes, respectively. HAP16 and HAP20 shared one derived mutation, consistent with an early split. Furthermore, given the subsequent accumulation of derived mutations not shared between HAP16 and HAP20, effectively no gene flow occurred between these two descendant populations, consistent with geographic separation. We therefore hypothesize that the common ancestor of these two clusters existed north of Cameroon among non-Bantu-speaking peoples in or around present-day Nigeria. From this common ancestral population, a group of migrants separated and traveled west and north to the area around present-day Senegal and the Gambia. These migrants could have taken a coastal or an inland route. The finding that the Senegal haplotype was the predominant haplotype in the sample of Mende from Sierra Leone is consistent with a coastal route. We do not have data to investigate an inland route, but we note that the frequency of the sickle allele is higher along the coast than inland.[53]

Classical haplotypes in eastern Arabia tend to have the Arabian/Indian designation, whereas those in western Arabia tend to have the Benin designation.[18,21] The Arabian/Indian haplotype has been hypothesized to have originated in either east Saudi Arabia or India.[5] Although our samples include only one predicted instance of the Arabian/Indian haplotype, the occurrence of this haplotype in the Luhya in Kenya and its clustering with the predominant haplotype found in Kenya and Uganda are consistent with the hypothesis that the Arabian/Indian haplotype originated in Africa and had an overseas migration route from eastern Africa to eastern Arabia and India.[13,19] In contrast, the absence of the Benin haplotype in the Luhya in Kenya and the Baganda in Uganda provides evidence against an overseas migration route from eastern Africa to western Arabia. Instead, the presence of the Benin haplotype in western Arabia is consistent with

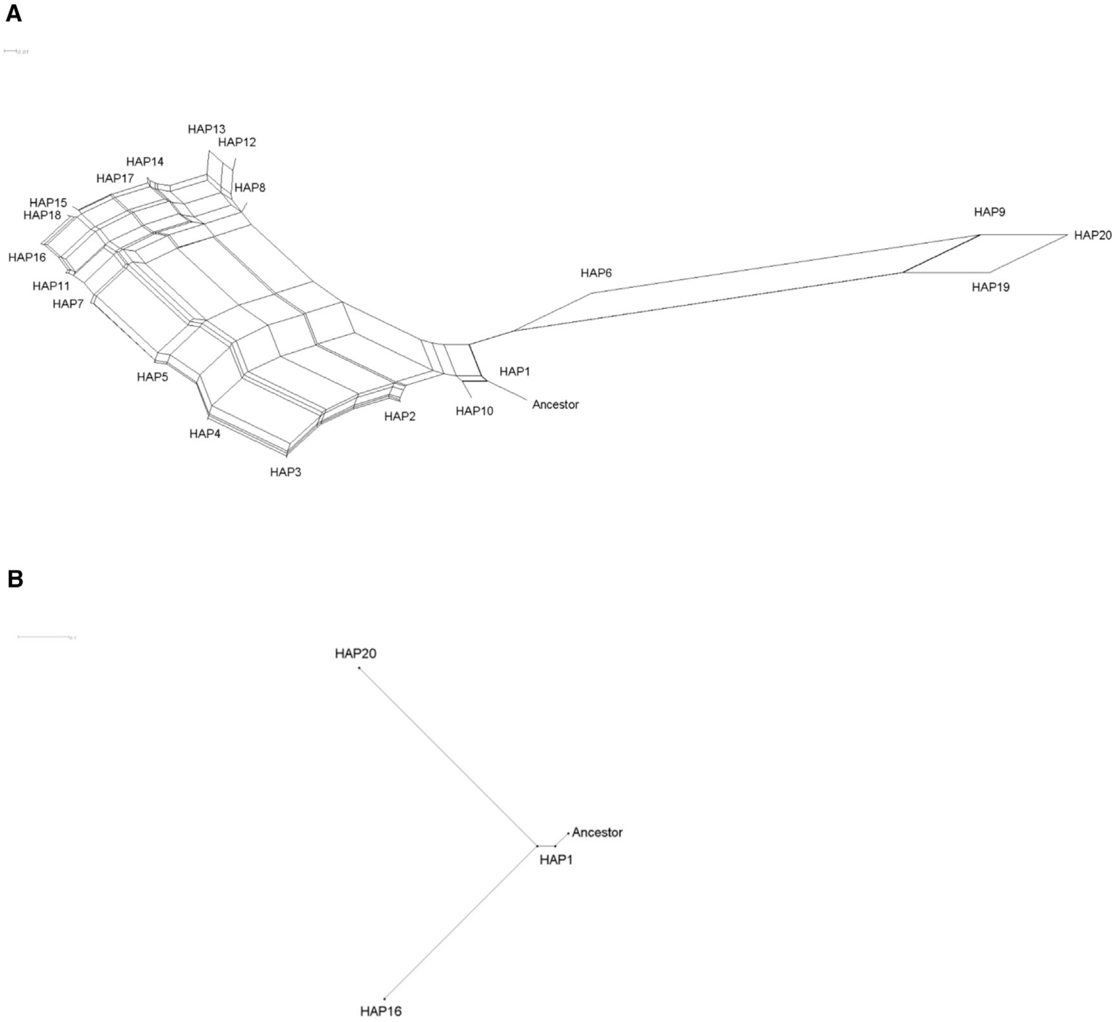

Figure 3. Split Decomposition Networks of Sickle-Carrying Haplotypes

(A) Network of 20 distinct sickle-carrying haplotypes, rooted by the ancestral haplotype. The haplotypes are defined in Table 4. The single branch leading from the ancestral root is the only branch to which the sickle mutation contributes, indicating a single origin of the sickle mutation. The scale bar represents 0.01 mutations/site.

(B) The subnetwork showing the ancestral root and the three modal haplotypes. This subnetwork emphasizes that HAP16 and HAP20 share a common ancestor and that this common ancestor is derived from HAP1. The scale bar represents 0.1 mutations/site.

an African origin and an overland migration route through northeast Africa.[33]

Together, our results suggest the following evolutionary history of the sickle mutation. The presence of African ancestry, an African patriline, and/or an African matriline in all sickle carriers, combined with the absence of Arabian or Indian ancestries in the five continental African samples in the 1000 Genomes Project,[33] supports an African origin of the sickle mutation. The sickle mutation occurred once approximately 7,300 years ago either in the Sahara or in west-central Africa. In an environment where malaria preexisted, balancing selection took approximately 2,400 years to drive the β^S allele to an equilibrium frequency of 12.0%. A population split occurred, possibly in the area around present-day Cameroon. Starting approximately 5,000 years ago, the Bantu expansions resulted in the spread of HAP1 to the CAR, the Great Lakes region, and South Africa. After acquiring one derived allele, the population split into one carrying a haplotype that evolved into HAP16 and another carrying a haplotype that evolved into HAP20. In both clusters, hitchhiking of the derived alleles with the β^S allele resulted in similar derived allele frequencies.

Our study includes sequence data from western, west-central, and eastern Africa but lacks comparable data from northern, central, and southern Africa. A haplotype has been identified among Sudanese individuals that was classified as atypical but might be a fifth African haplotype.[54] It is possible that sampling additional populations could provide more evidence regarding when, where, and in whom the sickle mutation arose.

Our findings support classification of sickle haplotypes based on three clusters. Notably, we found that the Senegal haplotype is substructured into two clusters, one containing only Senegal haplotypes and one containing Benin and Senegal haplotypes. This substructuring of haplotypes might have confounded previous assessments of clinical phenotype or disease severity. At the same time, the identification of markers linked to the sickle allele provides an opportunity to investigate possible cis-effects, such as differential expression of *HBD*, *HBE1*, and *HBG2*, on disease severity.

Supplemental Data

Supplemental Data include one figure and two tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.02.003.

Acknowledgments

We acknowledge the assistance of Neil Hanchard with the restriction site data. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official view of the National Institutes of Health. This research was supported by the Intramural Research Program of the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is supported by the National Human Genome Research Institute, the National Institute of Diabetes and Digestive and Kidney Diseases, the Center for Information Technology, and the Office of the Director at the National Institutes of Health (1ZIAHG200362).

Received: September 12, 2017
Accepted: February 2, 2018
Published: March 8, 2018

Web Resources

1000 Genomes Project, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

African Genome Variation Project, https://www.ebi.ac.uk/ega/studies/EGAS00001000959

Ensembl, http://www.ensembl.org/Homo_sapiens/Info/Index

E YTree, https://www.yfull.com/tree/E

Qatar sequence data, Sequence Read Archive, https://www.ncbi.nlm.nih.gov/sra/?term=SRP060765

OMIM, http://www.omim.org

References

1. Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., et al. (2001). Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. Science *293*, 455–462.
2. Carter, R., and Mendis, K.N. (2002). Evolutionary and historical aspects of the burden of malaria. Clin. Microbiol. Rev. *15*, 564–594.
3. Antonarakis, S.E., Boehm, C.D., Serjeant, G.R., Theisen, C.E., Dover, G.J., and Kazazian, H.H., Jr. (1984). Origin of the β ^{S}-globin gene in Blacks: the contribution of recurrent mutation or gene conversion or both. Proc. Natl. Acad. Sci. USA *81*, 853–856.
4. Pagnier, J., Mears, J.G., Dunda-Belkhodja, O., Schaefer-Rego, K.E., Beldjord, C., Nagel, R.L., and Labie, D. (1984). Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc. Natl. Acad. Sci. USA *81*, 1771–1773.
5. Kulozik, A.E., Wainscoat, J.S., Serjeant, G.R., Kar, B.C., Al-Awamy, B., Essan, G.J., Falusi, A.G., Haque, S.K., Hilali, A.M., Kate, S., et al. (1986). Geographical survey of β ^{S}-globin gene haplotypes: evidence for an independent Asian origin of the sickle-cell mutation. Am. J. Hum. Genet. *39*, 239–244.
6. Chebloune, Y., Pagnier, J., Trabuchet, G., Faure, C., Verdier, G., Labie, D., and Nigon, V. (1988). Structural analysis of the 5′ flanking region of the β-globin gene in African sickle cell anemia patients: further evidence for three origins of the sickle cell mutation in Africa. Proc. Natl. Acad. Sci. USA *85*, 4431–4435.
7. Lapouméroulie, C., Dunda, O., Ducrocq, R., Trabuchet, G., Mony-Lobé, M., Bodo, J.M., Carnevale, P., Labie, D., Elion, J., and Krishnamoorthy, R. (1992). A novel sickle cell mutation of yet another origin in Africa: the Cameroon type. Hum. Genet. *89*, 333–337.
8. Hanchard, N., Elzein, A., Trafford, C., Rockett, K., Pinder, M., Jallow, M., Harding, R., Kwiatkowski, D., and McKenzie, C. (2007). Classical sickle beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. BMC Genet. *8*, 52.
9. Bhagat, S., Patra, P.K., and Thakur, A.S. (2013). Fetal haemoglobin and β-globin gene cluster haplotypes among sickle cell patients in Chhattisgarh. J. Clin. Diagn. Res. *7*, 269–272.
10. Kan, Y.W., and Dozy, A.M. (1978). Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. Proc. Natl. Acad. Sci. USA *75*, 5631–5635.
11. Kan, Y.W., and Dozy, A.M. (1978). Antenatal diagnosis of sickle-cell anaemia by D.N.A. analysis of amniotic-fluid cells. Lancet *2*, 910–912.
12. Kurnit, D.M. (1979). Evolution of sickle variant gene. Lancet *1*, 104.
13. Mears, J.G., Lachman, H.M., Cabannes, R., Amegnizin, K.P., Labie, D., and Nagel, R.L. (1981). Sickle gene. Its origin and diffusion from West Africa. J. Clin. Invest. *68*, 606–610.
14. Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R.M., Clegg, J.B., Langaney, A., and Excoffier, L. (2002). Molecular analysis of the β-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the β^S Senegal mutation. Am. J. Hum. Genet. *70*, 207–223.
15. Solomon, E., and Bodmer, W.F. (1979). Evolution of sickle variant gene. Lancet *1*, 923.
16. Stine, O.C., Dover, G.J., Zhu, D., and Smith, K.D. (1992). The evolution of two west African populations. J. Mol. Evol. *34*, 336–344.
17. Flint, J., Harding, R.M., Clegg, J.B., and Boyce, A.J. (1993). Why are some genetic diseases common? Distinguishing selection from other processes by molecular analysis of globin gene variants. Hum. Genet. *91*, 91–117.
18. Ngo Bitoungui, V.J., Pule, G.D., Hanchard, N., Ngogang, J., and Wonkam, A. (2015). Beta-globin gene haplotypes among Cameroonians and review of the global distribution: is there a case for a single sickle mutation origin in Africa? OMICS *19*, 171–179.
19. Gelpi, A.P. (1973). Migrant populations and the diffusion of the sickle-cell gene. Ann. Intern. Med. *79*, 258–264.
20. Lehmann, H. (1954). Distribution of the sickle cell gene : A new light on the origin of the East Africans. Eugen. Rev. *46*, 101–121.
21. Livingstone, F.B. (1989). Who gave whom hemoglobin S: The use of restriction site haplotype variation for the

interpretation of the evolution of the β^S -globin gene. Am. J. Hum. Biol. *1*, 289–302.

22. Holloway, K., Lawson, V.E., and Jeffreys, A.J. (2006). Allelic recombination and *de novo* deletions in sperm in the human β-globin gene region. Hum. Mol. Genet. *15*, 1099–1111.
23. Srinivas, R., Dunda, O., Krishnamoorthy, R., Fabry, M.E., Georges, A., Labie, D., and Nagel, R.L. (1988). Atypical haplotypes linked to the β^S gene in Africa are likely to be the product of recombination. Am. J. Hematol. *29*, 60–62.
24. Papadakis, M.N., and Patrinos, G.P. (1999). Contribution of gene conversion in the evolution of the human β-like globin gene family. Hum. Genet. *104*, 117–125.
25. The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
26. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature *517*, 327–332.
27. Rodriguez-Flores, J.L., Fakhro, K., Agosto-Perez, F., Ramstetter, M.D., Arbiza, L., Vincent, T.L., Robay, A., Malek, J.A., Suhre, K., Chouchane, L., et al. (2016). Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. Genome Res. *26*, 151–162.
28. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
29. Bergsma, W. (2013). A bias-correction for Cramér's *V* and Tschuprow's *T*. J. Korean Stat. Soc. *42*, 323–328.
30. Jostins, L., Xu, Y., McCarthy, S., Ayub, Q., Durbin, R., Barrett, J., and Tyler-Smith, C. (2014). YFitter: maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data. arXiv, arXiv: 1407.7988, https://arxiv.org/abs/1407.7988.
31. Vianello, D., Sevini, F., Castellani, G., Lomartire, L., Capri, M., and Franceschi, C. (2013). HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment. Hum. Mutat. *34*, 1189–1194.
32. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
33. Baker, J.L., Rotimi, C.N., and Shriner, D. (2017). Human ancestry correlates with language and reveals that race is not an objective genomic classifier. Sci. Rep. *7*, 1572.
34. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A.; and 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
35. Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. *23*, 254–267.
36. Rasmussen, M.D., Hubisz, M.J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. PLoS Genet. *10*, e1004342.
37. R Core Team (2015). R: a language and environment for statistical computing (R Foundation for Statistical Computing).
38. Fenner, J.N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. *128*, 415–423.
39. Moorjani, P., Sankararaman, S., Fu, Q., Przeworski, M., Patterson, N., and Reich, D. (2016). A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. Proc. Natl. Acad. Sci. USA *113*, 5652–5657.
40. Shaikho, E.M., Farrell, J.J., Alsultan, A., Qutub, H., Al-Ali, A.K., Figueiredo, M.S., Chui, D.H.K., Farrer, L.A., Murphy, G.J., Mostoslavsky, G., et al. (2017). A phased SNP-based classification of sickle cell anemia *HBB* haplotypes. BMC Genomics *18*, 608.
41. Piel, F.B., Steinberg, M.H., and Rees, D.C. (2017). Sickle cell disease. N. Engl. J. Med. *376*, 1561–1573.
42. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40*, D930–D934.
43. Sallares, R., Bouwman, A., and Anderung, C. (2004). The spread of malaria to Southern Europe in antiquity: new approaches to old problems. Med. Hist. *48*, 311–328.
44. Institute of Medicine (US) Committee on the Economics of Antimalarial Drugs (2004). A Brief History of Malaria. In Saving Lives, Buying Time: Economics of Malaria Drugs in an Age of Resistance, K.J. Arrow, C. Panosian, and H. Gelband, eds. (National Academies Press), pp. 125–135.
45. Marin, A., Cerutti, N., and Massa, E.R. (1999). Use of the preamplification refractory mutation system (ARMS) in the study of HbS in predynastic Egyptian remains. Boll. Soc. Ital. Biol. Sper. *75*, 27–30.
46. Maat, G.J.R. (1993). Bone preservation, decay and its related conditions in ancient human bones from Kuwait. Int. J. Osteoarchaeol. *3*, 77–86.
47. Konotey-Ahulu, F.I.D. (1974). The sickle cell diseases. Clinical manifestations including the "sickle crisis". Arch. Intern. Med. *133*, 611–619.
48. Ehret, C. (2001). Bantu Expansions: re-envisioning a central problem of early African history. Int. J. Afr. Hist. Stud. *34*, 5–41.
49. Trombetta, B., D'Atanasio, E., Massaia, A., Ippoliti, M., Coppa, A., Candilio, F., Coia, V., Russo, G., Dugoujon, J.-M., Moral, P., et al. (2015). Phylogeographic refinement and large scale genotyping of human Y chromosome haplogroup E provide new insights into the dispersal of early pastoralists in the African continent. Genome Biol. Evol. *7*, 1940–1950.
50. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.
51. Castañeda, I.S., Mulitza, S., Schefuss, E., Lopes dos Santos, R.A., Sinninghe Damsté, J.S., and Schouten, S. (2009). Wet phases in the Sahara/Sahel region and human migration patterns in North Africa. Proc. Natl. Acad. Sci. USA *106*, 20159–20163.
52. Wainscoat, J.S. (1987). The origin of mutant β-globin genes in human populations. Acta Haematol. *78*, 154–158.
53. Piel, F.B., Patil, A.P., Howes, R.E., Nyangiri, O.A., Gething, P.W., Williams, T.N., Weatherall, D.J., and Hay, S.I. (2010). Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat. Commun. *1*, 104.
54. Elderdery, A.Y., Mills, J., Mohamed, B.A., Cooper, A.J., Mohammed, A.O., Eltieb, N., and Old, J. (2012). Molecular analysis of the β-globin gene cluster haplotypes in a Sudanese population with sickle cell anaemia. Int. J. Lab. Hematol. *34*, 262–266.