# Whole genome sequence-based haplotypes reveal single origin of the sickle allele during the Holocene Wet Phase

Daniel Shriner, Charles N. Rotimi*

Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, Maryland, 20892 USA

*Correspondence: rotimic@mail.nih.gov

ABSTRACT

Five classical designations of sickle haplotypes are based on the presence/absence of restriction sites and named after ethnic groups or geographic regions from which patients originated. Each haplotype is thought to represent an independent occurrence of the sickle mutation. We investigated the origins of the sickle mutation using whole genome sequence data. We identified 156 carriers from the 1000 Genomes Project, the African Genome Variation Project, and Qatar. We defined a new haplotypic classification 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/Bantu, Cameroon, and Arabian/Indian designations. Other haplotypes were derived from this haplotype and fell into two clusters, one comprised of haplotypes correlated with the Senegal designation and the other comprised of haplotypes correlated with both the Benin and Senegal designations. The near-exclusive presence of the original sickle haplotype in the Central African Republic, Kenya, Uganda, and South Africa is consistent with this haplotype predating the Bantu Expansion. 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 an age of the sickle mutation of 259 generations, corresponding to 7,300 years and 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. Many such variants are thought to have evolved in the last 10,000 years.[1; 2] In particular, the sickle allele $\beta^S$ in the beta globin gene *HBB* is a polymorphism under balancing selection due to recessive lethality and heterozygote advantage. The chromosomal background of the $\beta^S$ allele has been classified based on 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 named after ethno-linguistic groups or geographic regions, *i.e.*, Arabian/Indian, Benin, Cameroon, Central African Republic/Bantu, and Senegal, as well as a sixth category for "atypical" haplotypes.[3-9]

Whether the $\beta^S$ allele has a recent or old origin has been debated since the development of restriction fragment length polymorphism data.[10; 11] According to the multicentric model, the origin of the $\beta^S$ 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 $\beta^S$ allele is old, anywhere from tens to hundreds of thousands of years, 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 using whole genome sequence data from the 1000 Genomes Project, the African Genome Variation Project, and Qatar. We identified a total of 156 sickle carriers. Using phased sequence data, we established a new haplotypic classification. 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.

## MATERIALS AND METHODS

### *Ethics Statement*

This project was excluded from IRB 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 the 1000 Genomes Project,[25] the African Genome Variation Project,[26] and Qatar.[27] Haplotypes were delimited by positions ±500 kb of rs334 and pairwise $r^2 \geq 0.2$ with rs334. All data were processed using VCFtools version 0.1.14.[28]

### *African Ancestry*

Y chromosome haplogroups were called using YFitter.[29] Mitochondrial DNA haplogroups were called using HaploFind.[30] Autosomal ancestry was analyzed using projection analysis in ADMIXTURE version 1.3,[31] using a global reference panel of 21 global ancestries.[32] 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 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, $s = \frac{p}{1-2p}$. For each of the five continental African samples in the 1000 Genomes Project Phase 3 release version 5a, we estimated the effective population size $N_e$ based on the heterozygosities of all single nucleotide polymorphisms (*i.e.*, diallelic, triallelic, and quadrallelic), assuming a mutation rate of $0.97\times10^{-8}$ mutations/site/generation.[33] We then took the harmonic mean of the five $N_e$ estimates. We simulated 1,000 generations under a combination of random genetic drift and balancing selection, assuming one initial copy of the mutant allele. We repeated this process 1,000 times.

*Phylogenetic network analysis*

We used SplitsTree version 4.13.1 to perform split decomposition analysis of haplotypes.[34]

*Inferring the ancestral recombination graph*

We used ARGweaver to infer the ancestral recombination graph.[35] 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. We set the effective population size to the value described in the *Balancing Selection* subsection, the mutation rate to $0.97\times10^{-8}$ mutations/generation/site,[33] and the recombination rate to $1.5\times10^{-8}$ recombinations/generation/site. We also investigated larger recombination rates of $1.7\times10^{-8}$, $2.0\times10^{-8}$, and $2.0\times10^{-7}$ recombinations/generation/site. We used the functions heidel.diag and geweke.diag in the coda library of R, version 3.2.3, to assess

convergence diagnostics based on the posterior distribution of the number of recombination events [36] (Figure S1). To convert generations into years, we assumed a generation interval of 28 years.[37; 38]

## RESULTS

### *Molecular Mapping of Restriction Sites*

We mapped 15 restriction sites, including the 7 canonical sites, to the reference human genome sequence. We identified 12 known single nucleotide polymorphisms that predict the presence or absence of 10 of these sites. Of the canonical sites, we predict 5' ε HincII using rs3834466, Gγ1 HindIII using rs2070972, Aγ1 HindIII using rs28440105, and 3' ψβ HincII using rs968857 (Table 1); similar results were obtained using rs3834466, rs28440105, rs10128556, and rs968857.[39] Correlation (measured via $r^2$) between these variants and rs334 is weak to nonexistent (Table 1).

### *Distributions of $\beta^S$ and the classical haplotypes*

In the 1000 Genomes Project, we identified 137 sickle carriers and 0 sickle homozygotes; we predicted the classical haplotypes for all 137 carriers (Table 2). The average sickle allele frequency was 12.0% and did not statistically differ among the five continental African samples ( $\chi^2_4 = 1.48$, $p = 0.830$). 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. Of the 54 E1b1a, 29 were E1b1a1a1f and 23 were E1b1a1a1g.

In the African Genome Variation Project, we identified 14 sickle carriers in the Baganda and 1 sickle carrier in the Zulu. We predicted that all 15 of these individuals carried the Central African Republic/Bantu haplotype (Table 2). In the Qatar sample, we identified 4 sickle carriers, all with insufficient information to predict the haplotypes.

*New classification of haplotypes based on linkage disequilibrium*

We defined haplotypes centered on rs334 in the 504 continental Africans from the 1000 Genomes Project. First, we extracted 18,402 sites within 500 kb of rs334 with any non-reference allele count of 1. Then, we recorded pairwise LD for phased, diallelic sites (Table S1). The largest value of $r^2$ with rs334 was 0.407 and there were 27 sites with $r^2 \geq 0.2$. Based on rs334 and these 27 sites, we observed 62 haplotypes, of which 18 carried the sickle allele at rs334; a 19$^{th}$ sickle haplotype was observed once in the ACB sample and a 20$^{th}$ sickle haplotype was observed once in the Baganda (Table 3). The most common haplotype carried the ancestral allele at all 28 sites and accounted for 68.5% of all haplotypes. Thirteen 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 Central African Republic/Bantu (Table 3). Additionally, the autosomal fraction of African ancestry/patrilines/matrilines for the four Qatari carriers were 0.251/L1/L0a2a2a, 0.114/E1b1b1c*/H13c1, 0.974/E1b1a1a1f1a1/L3h1a2a1, and 0.078/NA/U6a2b1, indicating the presence of African ancestry in all four individuals. The three most common haplotypes (HAP1, HAP16, and HAP20) correlated primarily with the Central African Republic/Bantu, Benin, and Senegal designations, respectively.

*Bioinformatic annotation*

Different haplotypes might be associated with different clinical phenotypes or disease severity.[40] Using Ensembl and HaploReg version 4.1,[41] we annotated each of the 27 sites in linkage disequilibrium with rs334 (Table S2). Possible modifiers include nine sites marked on histones as promoters or enhancers and three sites bound by proteins. In addition, rs1039215 is correlated with gene expression, most strongly with *HBG2* (Table S3). HAP1 (correlated with

the Central African Republic/Bantu designation) and HAP16 (correlated with the Benin designation) differ by 13 sites, including rs73402608 (histone enhancer marks and bound protein) and rs1039215 (gene expression).

*Balancing selection*

We modeled balancing selection assuming that the relative fitness of the $\beta^A/\beta^A$ homozygote was 1, the relative fitness of the $\beta^S/\beta^S$ homozygote was 0, and the relative fitness of the $\beta^A/\beta^S$ heterozygote was $1+s$. Based on the 504 continental Africans from the 1000 Genomes Project, we estimated that $s = 0.158$, in agreement with previous estimates.[14] Next, we modeled random genetic drift plus balancing selection to estimate how many generations it would take for an equilibrium frequency of 12.0% to be reached, assuming a single initial copy and an effective population size $N_e = 25542$. 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. We stress that this value is not the age of the sickle mutation nor the age since the onset of balancing selection, but 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 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 2 generations.

*Phylogenetic network analysis*

We used split decomposition analysis to infer the phylogeny of the 20 sickle haplotypes, rooted by the ancestral haplotype (Figure 1). The network revealed that the sickle mutation occurred once in the background of the ancestral haplotype and gave rise to HAP1, associated

predominantly with the Central African Republic/Bantu designation. Two clusters were derived from this haplotype. One cluster contained HAP6, HAP9, HAP19, and HAP20, all associated with the Senegal designation. The other cluster contained haplotypes associated with both the Benin and Senegal designations.

*The ancestral recombination graph*

Using coalescent theory, we sampled the posterior distribution of the ancestral recombination graph using the 1,008 haplotypes, including 121 sickle haplotypes, from the 504 continental Africans from the 1000 Genomes Project. Conditional on this distribution, we estimated the age of the sickle mutation as 259 (95% confidence interval $[123, 395]$) generations. Recombination rates of $1.7\times10^{-8}$ recombinations/generation/site and larger yielded increasing numbers of incompatible sites.

## DISCUSSION

There are two models of the origins of the sickle allele. The multicentric model posits five independent occurrences of the same mutation in the last few thousand years. The unicentric model posits a single occurrence and an older age. We used whole genome sequence data to provide novel insight into this issue. Using a new 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.

The earliest recorded cases of malaria were ~5,000 years ago.[2; 42; 43] The first recorded cases of sickle were during the Hellenistic period, 2,130 years before present, in the Persian Gulf [44] and in 1670 AD in Ghana.[45] Based on these limited data, it is historically plausible that malaria preceded the sickle mutation, consistent with our simulations of balancing selection showing that the sickle allele would have been lost almost immediately without a heterozygote advantage (assuming recessive lethality). The simulations of balancing selection also indicated that it took ~2,400 years for equilibrium to be reached. This time provides a lower bound on the age of the sickle mutation, since we do not know how long the equilibrium state has been maintained.

The Bantu Expansion started ~5,000 years ago.[46] Our results imply that the sickle allele arose prior to the Bantu Expansion, consistent with the exclusive presence of the Central African Republic/Bantu haplotype in the Baganda and the Zulu. The Y chromosome haplogroups E1b1a1a1f and E1b1a1a1g, defined by L485 and U175 respectively, arose between 8,100 and 11,000 years before present. Our estimated age of the sickle mutation of ~7,300 years is consistent with a population in which these two sibling haplogroups co-circulated. Furthermore, our results place the origin of the sickle mutation 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.[47] Additionally, classical haplotypes in Western Arabia tend to have the Benin designation whereas those in Eastern Arabia tend to have the Arabian/Indian designation.[18; 21] Although our sample includes 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 suggest an Eastern African waypoint. The statistical presence of African ancestry in all sickle carriers, combined with the statistical absence of Arabian or Indian ancestries in the five continental African samples in the 1000 Genomes Project,[32] further supports an African origin of the sickle mutation.

We defined a new haplotypic classification based on phased sequence data. In contrast, the classical designations are based on restriction sites. By molecularly mapping restriction sites to the sequence data, we found that the restriction sites correlate poorly with rs334, such that none of the canonical sites was included in our sequence-based haplotypes. This result implies that the classical designations are not based on bona fide haplotypes. Our findings support a new classification based on three clusters. Notably, we found that the Senegal designation is substructured into two clusters, one of which shared with the Benin designation. This substructuring of haplotypes may have confounded previous assessments of clinical phenotype or disease severity.

## SUPPLEMENTAL DATA

Supplemental data include one figure and three tables.

## COMPETING INTERESTS

The authors declare no competing interests.

## ACKNOWLEDGEMENTS

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).

## WEB RESOURCES

Ensembl, http://www.ensembl.org/Homo_sapiens/Info/Index; E YTree, https://www.yfull.com/tree/E

# 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 $\beta^S$-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 $\beta^S$-globin gene haplotypes: evidence for an independent Asian origin of the sickle-cell mutation. Am. J. Hum. Genet. *39*, 239-244.

6. Cheboline, 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 $\beta^S$ 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 $\beta^S$ -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 $\beta^S$ 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. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158.

29. 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, 1407.7988.

30. 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.

31. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655-1664.

32. 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.

33. The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

34. Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. *23*, 254-267.

35. Rasmussen, M.D., Hubisz, M.J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. PLOS Genet. *10*, e1004342.

36. R Core Team. (2015). R: a language and environment for statistical computing. (Vienna, Austria: R Foundation for Statistical Computing).

37. 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.

38. 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.

39. 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.

40. Piel, F.B., Steinberg, M.H., and Rees, D.C. (2017). Sickle cell disease. N. Engl. J. Med. *376*, 1561-1573.

41. 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.

42. 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.

43. 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. (Washington (DC), National Academies Press (US)).

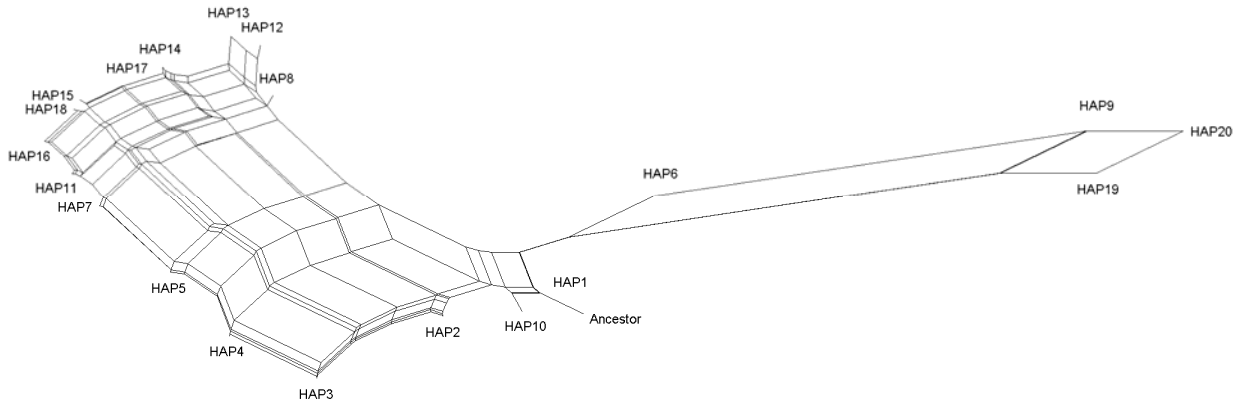
44. Maat, G.J.R. (1993). Bone preservation, decay and its related conditions in ancient human bones from Kuwait. Int. J. Osteoarchaeol. *3*, 77-86.

45. Konotey-Ahulu, F.I.D. (1974). The sickle cell diseases: clinical manifestations including the "sickle crisis". Arch. Intern. Med. *133*, 611-619.

46. Ehret, C. (2001). Bantu Expansions: re-envisioning a central problem of early African history. Int. J. Afr. Hist. Stud. *34*, 5-41.

47. 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.

1 FIGURE LEGEND




2

3 Figure 1. Split decomposition network of 20 distinct sickle-carrying haplotypes, rooted by the

4 ancestral haplotype.

Table 1. Molecular characterization of the classical sickle designations.

| Site | Sequence [a] | Range | rsID | Position (hg19) | Senegal | Benin | CAR/ Bantu | Cameroon | Arabian/ Indian | $r^{2}$ [b] | D' [b] | Ancestral | Status | Derived | Status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3' ψβ HincII | GTTGAC | 5260457-5260462 | rs968857 | 5260458 | + | + | - | + | + | 0.000 | -0.104 | T | + | C | - |
| Aγ1 HindIII | AAGCTT | 5269799-5269804 | rs28440105 | 5269799 | - | - | - | + | - | 0.016 | 0.930 | C | - | A | + |
| Gγ1 HindIII | AAGCTT | 5274717-5274722 | rs2070972 | 5274717 | + | - | + | + | + | 0.003 | 0.094 | C | - | A | + |
| 5' ε HincII | GTTGAC | 5291563-5291567 | rs3834466 | 5291563-5291564 | - | - | - | - | + | 0.020 | -0.853 | G | - | GT | + |

[a] Red indicates the polymorphic position.

[b] Pairwise linkage disequilibrium values are shown with respect to rs334.

Table 2. Distribution of classical sickle designations.

| Sample [a] | Arabian/Indian | Benin | Cameroon | CAR/Bantu | 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: 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”.

Table 3. Distribution of sickle haplotypes under a sequence-based classification scheme.

| Name | Haplotype [a] | Arabian/Indian | Benin | Cameroon | CAR/Bantu | Senegal | Atypical |
|---|---|---|---|---|---|---|---|
| Ancestor | 00000000000000000000000000010 | NA | NA | NA | NA | NA | NA |
| HAP1 | 00000000000000000000010000010 | 1 | 0 | 2 | 37 | 0 | 1 |
| HAP2 | 00000000000000000000010110010 | 0 | 4 | 0 | 0 | 0 | 0 |
| HAP3 | 00000000000000000000010110101 | 0 | 0 | 0 | 0 | 1 | 0 |
| HAP4 | 00000000000000000001110110101 | 0 | 2 | 0 | 0 | 0 | 0 |
| HAP5 | 00000000000000000111110110101 | 0 | 5 | 0 | 0 | 0 | 0 |
| HAP6 | 00000000000000001100011001010 | 0 | 0 | 0 | 0 | 3 | 0 |
| HAP7 | 00000001000001001111110110101 | 0 | 1 | 0 | 0 | 0 | 0 |
| HAP8 | 00000001000001001111110110110 | 0 | 0 | 0 | 0 | 1 | 0 |
| HAP9 | 00001110111110111100011001010 | 0 | 0 | 0 | 0 | 1 | 0 |
| HAP10 | 00010000000000000000010000010 | 0 | 0 | 0 | 1 | 0 | 0 |
| HAP11 | 00010001000001001111110110101 | 0 | 2 | 0 | 0 | 1 | 0 |
| HAP12 | 00110001000001001011100000010 | 0 | 0 | 0 | 0 | 1 | 0 |
| HAP13 | 00110001000001001111110000010 | 0 | 1 | 0 | 0 | 0 | 1 |
| HAP14 | 00110001000001001111110110010 | 0 | 0 | 0 | 0 | 1 | 0 |
| HAP15 | 00110001000001001111110110100 | 0 | 2 | 0 | 0 | 1 | 0 |
| HAP16 | 00110001000001001111110110101 | 0 | 23 | 0 | 0 | 7 | 0 |
| HAP17 | 00110001000001001111110110110 | 0 | 6 | 0 | 0 | 5 | 0 |
| HAP18 | 00110001000001001111110110111 | 0 | 1 | 0 | 0 | 0 | 0 |
| HAP19 | 11001110111110111000010000010 | 0 | 0 | 0 | 0 | 3 | 1 |
| HAP20 | 11001110111110111000011001010 | 0 | 0 | 0 | 0 | 36 | 0 |

[a] 0 indicates the reference allele and 1 indicates the alternate allele, following the coding scheme in the 1000 Genomes Project vcf files.