REPORT

Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians

Luca Pagani,[1,2,3,*] Stephan Schiffels,[1] Deepti Gurdasani,[1] Petr Danecek,[1] Aylwyn Scally,[4] Yuan Chen,[1] Yali Xue,[1] Marc Haber,[1,5] Rosemary Ekong,[6] Tamiru Oljira,[7] Ephrem Mekonnen,[7] Donata Luiselli,[3] Neil Bradman,[8] Endashaw Bekele,[7] Pierre Zalloua,[5,9] Richard Durbin,[1] Toomas Kivisild,[2] and Chris Tyler-Smith[1]

The predominantly African origin of all modern human populations is well established, but the route taken out of Africa is still unclear. Two alternative routes, via Egypt and Sinai or across the Bab el Mandeb strait into Arabia, have traditionally been proposed as feasible gateways in light of geographic, paleoclimatic, archaeological, and genetic evidence. Distinguishing among these alternatives has been difficult. We generated 225 whole-genome sequences (225 at 8× depth, of which 8 were increased to 30×; Illumina HiSeq 2000) from six modern Northeast African populations (100 Egyptians and five Ethiopian populations each represented by 25 individuals). West Eurasian components were masked out, and the remaining African haplotypes were compared with a panel of sub-Saharan African and non-African genomes. We showed that masked Northeast African haplotypes overall were more similar to non-African haplotypes and more frequently present outside Africa than were any sets of haplotypes derived from a West African population. Furthermore, the masked Egyptian haplotypes showed these properties more markedly than the masked Ethiopian haplotypes, pointing to Egypt as the more likely gateway in the exodus to the rest of the world. Using five Ethiopian and three Egyptian high-coverage masked genomes and the multiple sequentially Markovian coalescent (MSMC) approach, we estimated the genetic split times of Egyptians and Ethiopians from non-African populations at 55,000 and 65,000 years ago, respectively, whereas that of West Africans was estimated to be 75,000 years ago. Both the haplotype and MSMC analyses thus suggest a predominant northern route out of Africa via Egypt.

The routes followed by fully modern humans as they expanded out of Africa 50,000–100,000 years ago into Eurasia have long been a central question of anthropology[1] and have important implications for understanding the evolutionary history of all non-African populations. So far, neither fossil and archaeological[2–4] nor genetic[5,6] evidence has been able to distinguish between an exit through Egypt and Sinai (northern route)[7] or one through Ethiopia, the Bab el Mandeb strait, and the Arabian Peninsula (southern route).[8–10] Genetic evidence has more often been interpreted as favoring a southern route,[5,6,9] although the Neandertal admixture present in all non-Africans[11] is more readily explained by a northern route given that Neandertal fossils are currently known from the Levant, but not from the southern part of the Arabian Peninsula.[12] Thus, the available evidence remains inconclusive. Information to discriminate between the northern and southern routes might still be present in Africa within the full genomes of the populations inhabiting modern Egypt and the Horn of Africa, and thus further investigation is warranted. However, although it might not be easy to extract this information because of the past and recent genetic introgression experienced by these populations,[13,14] full sequences of Northeast African genomes would provide the best starting point for these and other analyses.

To improve our understanding of the African gene pool that might have been ancestral to the out-of-Africa (OOA) dispersal, we sequenced the genomes of a random sample of 100 Egyptians and 125 individuals from five Ethiopian populations (25 each from Amhara, Oromo, Ethiopian Somali, Wolayta, and Gumuz) to an average depth of 8× by using an Illumina HiSeq 2000, and we analyzed these data within the context of similar data generated by the 1000 Genomes Project.[15] Sample collection, export, and analysis were approved by University College London research ethics committee 0489/002, Ethiopian Ministry of Science and Technology approval no. 310/538/04, and Lebanese American University institutional review board SMPZ121307-2 (see Supplemental Data for additional information). The overall genetic landscape emerging from the sequencing data (Table S1) refines current knowledge of the high diversity in the Ethiopian region. Sequence data avoid the effect of ascertainment bias that one encounters when dealing with SNP arrays from the same populations (Figure S1). If the

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK; [2]Department of Archaeology and Anthropology, University of Cambridge, Cambridge CB2 1QH, UK; [3]Department of Biological, Geological, and Environmental Sciences, University of Bologna, 40126 Bologna, Italy; [4]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK; [5]The Lebanese American University, Chouran, Beirut 1102 2801, Lebanon; [6]Department of Genetics, Evolution, and Environment, University College London, London WC1E 6BT, UK; [7]University of Addis Ababa and Center of Human Genetic Diversity, PO Box 1176, Ethiopia; [8]Henry Stewart Group, 28/30 Little Russell Street, London WC1A 2HN, UK; [9]Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
*Correspondence: lp.lucapagani@gmail.com
http://dx.doi.org/10.1016/j.ajhg.2015.04.019. ©2015 The Authors
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

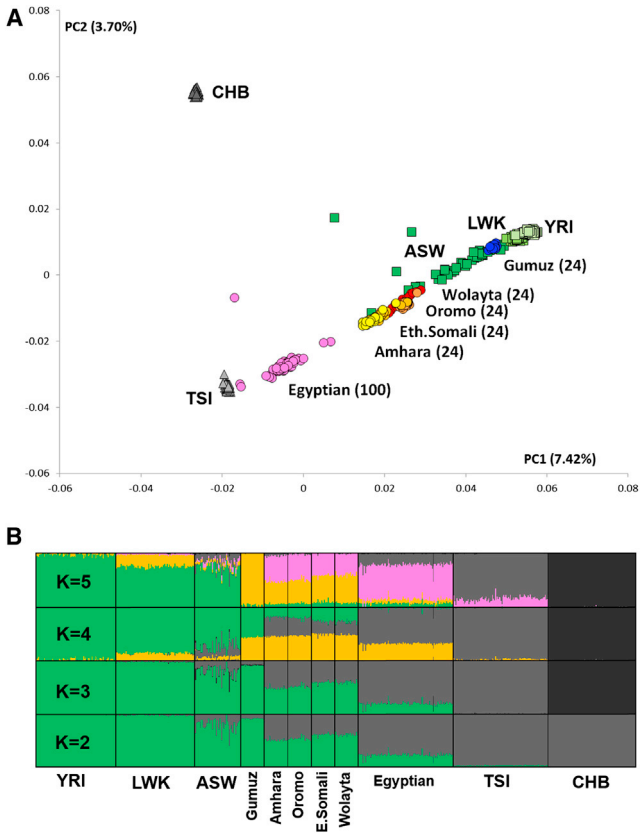

Figure 1. PCA and ADMIXTURE Analysis

PCA (A) and ADMIXTURE analysis (B) of the newly sequenced samples (Egyptian, pink; Amhara, yellow; Oromo and Ethiopian Somali, light orange; Wolayta, red; and Gumuz, blue) and a subset of 1000 Genomes samples (CHB, dark gray; TSI, light gray; ASW [African ancestry in Southwest USA], green; and LWK [Luhya in Webuye, Kenya] and YRI, light green). ADMIXTURE was run with different values of K ($K = 5$ was the smallest cross-validation error). The top ADMIXTURE plot shows five ancestral components tentatively describable as West African (green), East African (orange), European (light gray), East Asian (dark gray), and putatively Middle Eastern (pink). The phased and imputed genotypes from the low-coverage sequences were processed with PLINK[19] for the removal of variants with a minor allele frequency $< 1\%$ (--maf 0.01 --geno 0.01) and pairwise linkage disequilibrium above 0.1 (--indep-pairwise 50 10 0.1). The pruned dataset was then analyzed by ADMIXTURE[17] with the --cv option for assessing the most plausible value of K and also by PCA.[18] The proportion of the total variance explained by each principal component is reported as a percentage next to each axis label.

northern route was the predominant path followed by the ancestors of the OOA populations, and modern African populations are representative of those at the time of the exit, Egyptians should be genetically more similar to modern non-Africans. Conversely, if the southern route was the main way out of Africa, Ethiopians should be closest to the OOA populations. However, extensive historical and genetic data show that recent gene flow has drastically influenced the genomes of present-day Egyptians and Ethiopians.[13,14,16] To minimize the confounding effect of this gene flow back to Africa while testing this hypothesis, we first identified and then masked the recent non-African ancestry in the Ethiopian and Egyptian genomes.

Using ADMIXTURE[17] and principal-component analysis (PCA)[18] (Figure 1A), we estimated the average proportion of non-African ancestry in the Egyptians to be 80% and dated the midpoint of the admixture event by using ALDER[20] to around 750 years ago (Table S2), consistent with the Islamic expansion and dates reported previously.[13,14] The Ethiopian populations showed, as expected, a more variable spectrum of genetic introgression (Figure 1B). Consistent with previous reports,[13] the Amhara and Oromo were shown to have around 50% of their genome derived from non-Africans, the introgressed proportion in the Somali and Wolayta amounted to 40%–30%, and the Gumuz showed negligible amounts of non-African admixture. The date of the midpoint of these admixture events was 2,500–3,000 years ago (Table S2), although one notable exception was the Oromo, who have shown evidence of multiple admixture events.[21] These conclusions are consistent with previous reports[13,21] and fit with linguistic records.[22] Furthermore, the distribution of maternal (mtDNA) and paternal (Ychr) lineages revealed sex-biased admixture patterns in Ethiopians (Figure S2), such that there was less male-mediated than female-mediated Middle Eastern backflow. The affinity of the Egyptian African component with the modern East and West African populations (green component in Figure 1B, $K = 5$) could be due to either a continuity of human presence in the area or recent gene flow from neighboring African regions resulting from demographic processes and slave trade over the last two millennia.[23]

In order to filter out, through masking, the Eurasian portion identified in this way, we phased the samples by using ShapeIT[24] and processed them with PCAdmix.[25] In the masking process, Europeans (CEU [Utah residents with ancestry from northern and western Europe from the CEPH collection])[15] were used as a proxy for the non-African component, and the Gumuz (the Ethiopian population showing minimal introgression) were used as a proxy for the African component. Pairwise F_{ST}[26] was calculated before and after the masking process (Table S3), highlighting the expected trend of increased distance of the admixed populations from non-Africans when we retained only their African component. After we excluded the Gumuz themselves from the subsequent analyses, we compared the African components of the masked Ethiopian and Egyptian genomes (hereafter referred to as the Ethiopian′ and Egyptian′ genomes, respectively) with a set of West African (YRI [Yoruba in Ibadan, Nigeria]) and OOA populations spanning Eurasia (East Asian CHB [Han Chinese in Beijing, China], European TSI [Toscani in Italia] and CEU [Figure 2], and South Asian GIH [Gujarati Indians in Houston, Texas] [Figure S6]) in order to look for a signature of the OOA migration. Such a signature was defined as a higher similarity between the Ethiopian′ or Egyptian′ genomes and the non-Africans than between the latter and

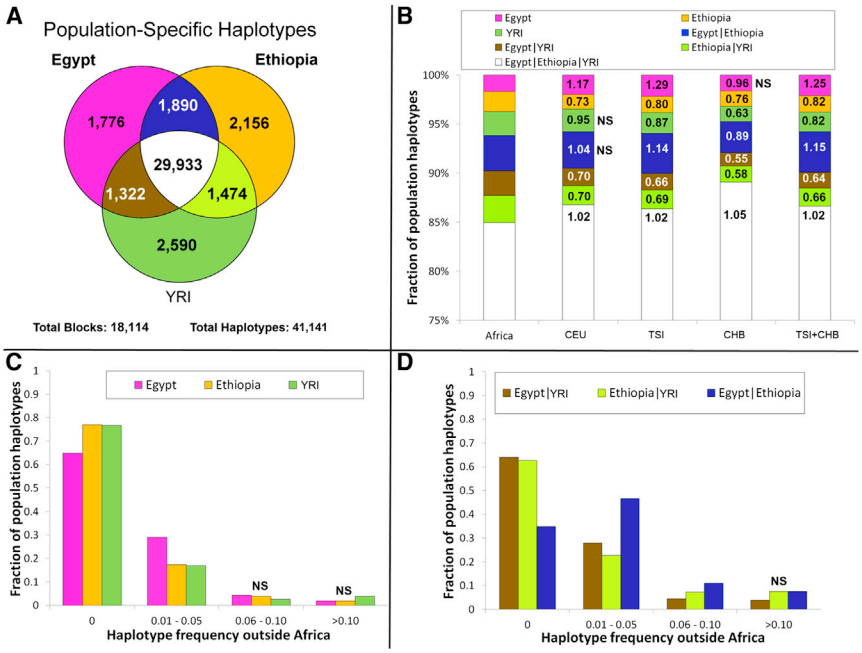

Figure 2. Haplotype Sharing between African and Non-African Populations

The 41,141 African haplotypes retrieved from 18,114 LD regions outside Africa were grouped according to the population of discovery (A). The haplotype composition of African and non-African (CHB + TSI) populations (B) showed more Egyptian′ (pink) and Egyptian′|Ethiopian′ (blue)-specific haplotypes in the OOA samples (relative increases from the general African population are provided for each colored section) than did the haplotype composition of the combined African populations. Non-significant (χ^2i) comparisons are labeled "NS." Of the haplotypes specific to a single African population, the Egyptian′ haplotypes (pink) showed the highest population frequency outside Africa (C), whereas the Egyptian′|Ethiopian′ haplotypes (blue) were the most frequent of those shared by two African populations (D). Bars not significantly different (tested with χ^2i) from the Egyptian′ (C) or Ethiopian′|Egyptian′ (D) ones are labeled "NS." The first bin in (C) and (D) shows the proportion of African haplotypes not present outside Africa.

the YRI. If we assume a stepwise differentiation out of Africa, and if the preferential route followed was the northern one, Egyptian′ samples should share the highest number of haplotypes with the Eurasian samples even after recent events of introgression are controlled for. Conversely, Ethiopian′ samples would show the highest haplotype sharing with the Eurasian samples if the southern route was preferentially followed during the OOA migration. We restricted this comparison to 18,114 genomic regions (spanning a total length of 7.1 Mb; Figure S5) containing haplotypes shared by Europeans and Asians because these were likely to predate the split between these populations. Given the broad occurrence of these regions outside Africa, we could rule out positive selection as a plausible driver of the observed linkage-disequilibrium (LD) pattern. We identified these regions by calculating LD blocks in a set of 457 non-African samples. We retrieved 41,141 haplotypes at these loci in the Egyptian′, Ethiopian′, or YRI samples (Figure 2A) and used them to estimate the genetic similarity between OOA populations CHB and TSI and each of the three African populations. 85% of the haplotypes were present in all three African populations and were discarded as non-informative. The remaining 15% of haplotypes were instead observed in only one or two African populations. For these haplotypes that could discriminate between the African populations, the combined CHB and TSI samples showed more Egyptian′-specific (1.25-fold, $p = 2 \times 10^{-6}$) and Ethiopian′- and Egyptian′-specific (hereafter Ethiopian′|Egyptian′-specific) (1.15-fold, $p = 9 \times 10^{-6}$) haplotypes than did any of the other African haplotype sets (Figure 2B). We further explored the observed enrichment of Egyptian′ haplotypes in the CHB and TSI samples by investigating the frequency of each class of haplotype in the combined CHB and TSI samples, and again, the frequencies of Egyptian′-specific and Egyptian′|Ethiopian′-specific haplotypes were highest (Figures 2C and 2D). The enrichment of Egyptian′ haplotypes in the genetic pool of the CHB and TSI samples points to a northern migration as the greater contributor to populations outside Africa.

This finding was robust to a wide range of potential artifacts stemming from uncertainties in the masking process (Figures S3, S4, and S6A; Table S4; note particularly the false-positive rate displayed in column 8) and was replicated in a South Asian population (GIH; Figure S6B). Furthermore, we showed with simulations that the error rate present in the masking process (Table S4) was unlikely to affect our findings (Figures S4 and S6). Even when we added a 10% misclassification error to the Ethiopians, Egyptians held as the African population showing the highest affinity to non-Africans. Alternative scenarios involving early back-to-Africa migrations[27] as the source of haplotype sharing between Egyptian′ and non-African samples were considered as sources of the observed pattern. However, such confounding backflow would need to have taken place prior to the split between East Asians and Europeans (ca. ~40,000 years ago) and, if this genetic component originated from the main OOA founding event, is likely to have been removed by the non-African masking procedure, which was designed for this purpose.

To provide an independent test of our finding, we analyzed three Egyptian and five Ethiopian high-coverage genomes with the multiple sequentially Markovian coalescent (MSMC) approach before and after masking and

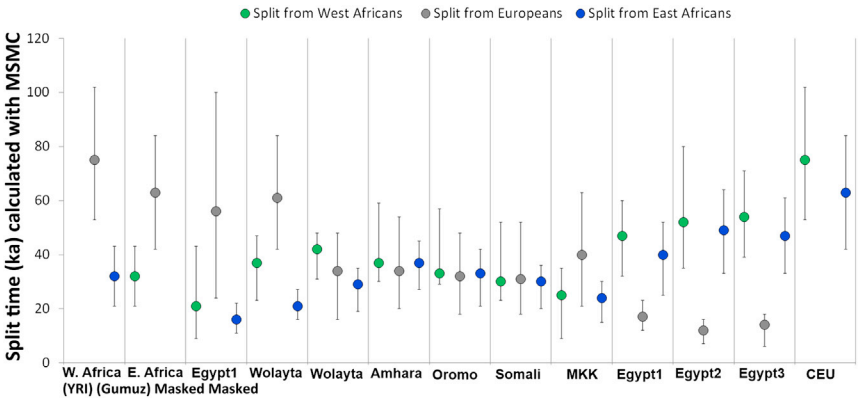

Figure 3. Inferred Split Times between Pairs of High-Coverage Genomes

MSMC-inferred genetic split times of a set of five Ethiopian, three Egyptian, one Maasai, one European (CEU), and one West African (YRI) randomly chosen genome from Europeans, West Africans, and East Africans (Gumuz). One Egyptian (Egypt1) and one Ethiopian (Wolayta) genome were analyzed also after their non-African component was masked out. The split time between two genomes is defined as the time when the cross-coalescence rate dropped to 50%. Cross-coalescence rates of 75% and 25% are shown by the top and bottom bars, respectively, providing references for the putative beginning and end, respectively, of the population split event. The space covered by each vertical line is therefore intended to provide a "time range" when the population split might have occurred, thus showing the split between populations as a slow rather than an instantaneous phenomenon.

compared them with a set of publicly available high-coverage genomes.[15,28] MSMC,[29] an extension of the PSMC[30] method to two or four genomes, estimates the split time between pairs of genomes. Consistent with their admixed nature, the split times of the non-masked Egyptians and the mixed Ethiopians from Europeans (CEU) and West Africans (YRI) were much closer to each other than to the same split times measured in the non-admixed Ethiopian population (Gumuz) (Figure 3; Figure S7). If we consider the genetic split between two populations as a process gradually occurring over thousands of years, two independent splits might show partial overlaps when their midpoints are less than a few thousand years apart. Keeping in mind this potential confounder, the Ethiopian′ and Egyptian′ genomes showed different patterns. In particular, the Egyptian′ genomes displayed a more recent split from both the West African (21,000 years ago) and the non-African (55,000 years ago) genomes than did the Ethiopian′ genomes (37,000 and 65,000 years ago, respectively). This suggests a higher similarity between non-African and Egyptian′ components than between non-African and Ethiopian′ components, which is consistent with the fact that Egypt is the last stop on the way out of Africa. Such split dates[21] also hint at a recent interaction between Egyptians and West Africans (Figure 3).

In conclusion, the analysis of Ethiopian′ and Egyptian′ whole-genome sequence data identifies modern Egyptians as the African population whose genome and haplotype frequency most closely resemble those of non-African populations. The fact that we could identify in Egyptians an African genomic component that is distinct from West and East African components further supports a minor degree of population continuity in Egypt since the OOA dispersal. These findings point to the northern route as the preferential direction taken out of Africa. In doing this, they resolve the puzzles of archaeological similarities and Neandertal admixture, which are readily accommodated by a northern-exit model, but not by a southern exit, and fit well with the recent discovery of human remains dating to around 55,000 years ago in Israel (close to the northern route).[31] Furthermore, the data generated here provide a better source of information for spatially explicit demographic models.[32,33] Our analysis does not address controversies about the timing and possible complexities of the expansion out of Africa and highlights the need for further analyses, ideally including ancient DNA, as well as Near Eastern and Papuan or Australian genomes representative of an early coastal expansion, to further resolve these issues.

Accession Numbers

The European Genome-phenome Archive (EGA) accession numbers for the Egyptian and Ethiopian sequences reported in this paper are EGA: EGAS00001000480 (Egyptian low coverage), EGAS00001000482 (Egyptian high coverage), EGAS00001000238 (Ethiopian low coverage), EGAS00001000237 (Ethiopian high coverage). SNPchip data (.bed, .bim, and .fam) and called genotype files (.vcf) are available from the corresponding author upon request.

Supplemental Data

Supplemental Data include seven figures and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.04.019.

Acknowledgments

The authors would like to acknowledge all the donors who kindly contributed samples to this study. L.P., S.S., D.G., P.D., Y.C., Y.X., M.H., R.D., and C.T.-S. were funded by Wellcome Trust grant 098051. T.K. was funded by European Research Council (ERC) Starting Grant FP7-261213, and D.L. was funded by ERC Advanced Grant FP7-295733. R.D. is a founder and non-executive director of Congenica. N.B. is the senior trustee, settlor, and principal donor of Melford Charitable Trust and the director and beneficial owner of the entire share capital of Cordell Homes, a company that provided the financial support for the collection of the Ethiopian samples analyzed in this paper.

Received: January 21, 2015
Accepted: April 29, 2015
Published: May 28, 2015

Web Resources

The URL for data presented in this paper is as follows:

European Genome-phenome Archive (EGA), https://www.ebi.ac.uk/ega/home

References

1. Jobling, M.A., Hollox, E., Hurles, M., Kivisild, T., and Tyler-Smith, C. (2013). Human evolutionary genetics (New York: Garland Science).
2. Armitage, S.J., Jasim, S.A., Marks, A.E., Parker, A.G., Usik, V.I., and Uerpmann, H.P. (2011). The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia. Science *331*, 453–456.
3. Petraglia, M.D., Haslam, M., Fuller, D.Q., Boivin, N., and Clarkson, C. (2010). Out of Africa: new hypotheses and evidence for the dispersal of Homo sapiens along the Indian Ocean rim. Ann. Hum. Biol. *37*, 288–311.
4. Stringer, C.B., Grün, R., Schwarcz, H.P., and Goldberg, P. (1989). ESR dates for the hominid burial site of Es Skhul in Israel. Nature *338*, 756–758.
5. Soares, P., Alshamali, F., Pereira, J.B., Fernandes, V., Silva, N.M., Afonso, C., Costa, M.D., Musilová, E., Macaulay, V., Richards, M.B., et al. (2012). The Expansion of mtDNA Haplogroup L3 within and out of Africa. Mol. Biol. Evol. *29*, 915–927.
6. Quintana-Murci, L., Semino, O., Bandelt, H.J., Passarino, G., McElreavey, K., and Santachiara-Benerecetti, A.S. (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat. Genet. *23*, 437–441.
7. Campbell, M.C., and Tishkoff, S.A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu. Rev. Genomics Hum. Genet. *9*, 403–433.
8. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The history and geography of human genes (Princeton, N.J.: Princeton University Press).
9. Cavalli-Sforza, L.L., Piazza, A., Menozzi, P., and Mountain, J. (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc. Natl. Acad. Sci. USA *85*, 6002–6006.
10. Lahr, M., and Foley, R. (1994). Multiple Dispersals and Modern Human Origin. Evol. Anthropol. *3*, 48–60.
11. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.
12. Krause, J., Orlando, L., Serre, D., Viola, B., Prüfer, K., Richards, M.P., Hublin, J.J., Hänni, C., Derevianko, A.P., and Pääbo, S. (2007). Neanderthals in central Asia and Siberia. Nature *449*, 902–904.
13. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., et al. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. *91*, 83–96.
14. Henn, B.M., Botigué, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlaoui-Zid, K., Zalloua, P.A., Moreno-Estrada, A., Bertranpetit, J., et al. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. *8*, e1002397.
15. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
16. Haber, M., Gauguier, D., Youhanna, S., Patterson, N., Moorjani, P., Botigué, L.R., Platt, D.E., Matisoo-Smith, E., Soria-Hernanz, D.F., Wells, R.S., et al. (2013). Genome-wide diversity in the levant reveals recent structuring by culture. PLoS Genet. *9*, e1003316.
17. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
18. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
19. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
20. Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. Genetics *193*, 1233–1254.
21. Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. Proc. Natl. Acad. Sci. USA *111*, 2632–2637.
22. Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C.J. (2009). Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc. Biol. Sci. *276*, 2703–2710.
23. Serjeant, G.R. (1997). Sickle-cell disease. Lancet *350*, 725–730.
24. Delaneau, O., Marchini, J., and Zagury, J.F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.
25. Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J.G., and Bustamante, C.D. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum. Biol. *84*, 343–364.
26. Wright, S. (1931). Evolution in Mendelian Populations. Genetics *16*, 97–159.
27. Hodgson, J.A., Mulligan, C.J., Al-Meeri, A., and Raaum, R.L. (2014). Early back-to-Africa migration into the Horn of Africa. PLoS Genet. *10*, e1004393.
28. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78–81.
29. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925.

30. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.
31. Hershkovitz, I., Marder, O., Ayalon, A., Bar-Matthews, M., Yasur, G., Boaretto, E., Caracuta, V., Alex, B., Frumkin, A., Goder-Goldberger, M., et al. (2015). Levantine cranium from Manot Cave (Israel) foreshadows the first European modern humans. Nature *520*, 216–219.
32. Eriksson, A., and Manica, A. (2012). Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. USA *109*, 13956–13960.
33. Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. *43*, 1031–1034.

The American Journal of Human Genetics

Supplemental Data

Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians

Luca Pagani, Stephan Schiffels, Deepti Gurdasani, Petr Danecek, Aylwyn Scally, Yuan Chen, Yali Xue, Marc Haber, Rosemary Ekong, Tamiru Oljira, Ephrem Mekonnen, Donata Luiselli, Neil Bradman, Endashaw Bekele, Pierre Zalloua, Richard Durbin, Toomas Kivisild, and Chris Tyler-Smith

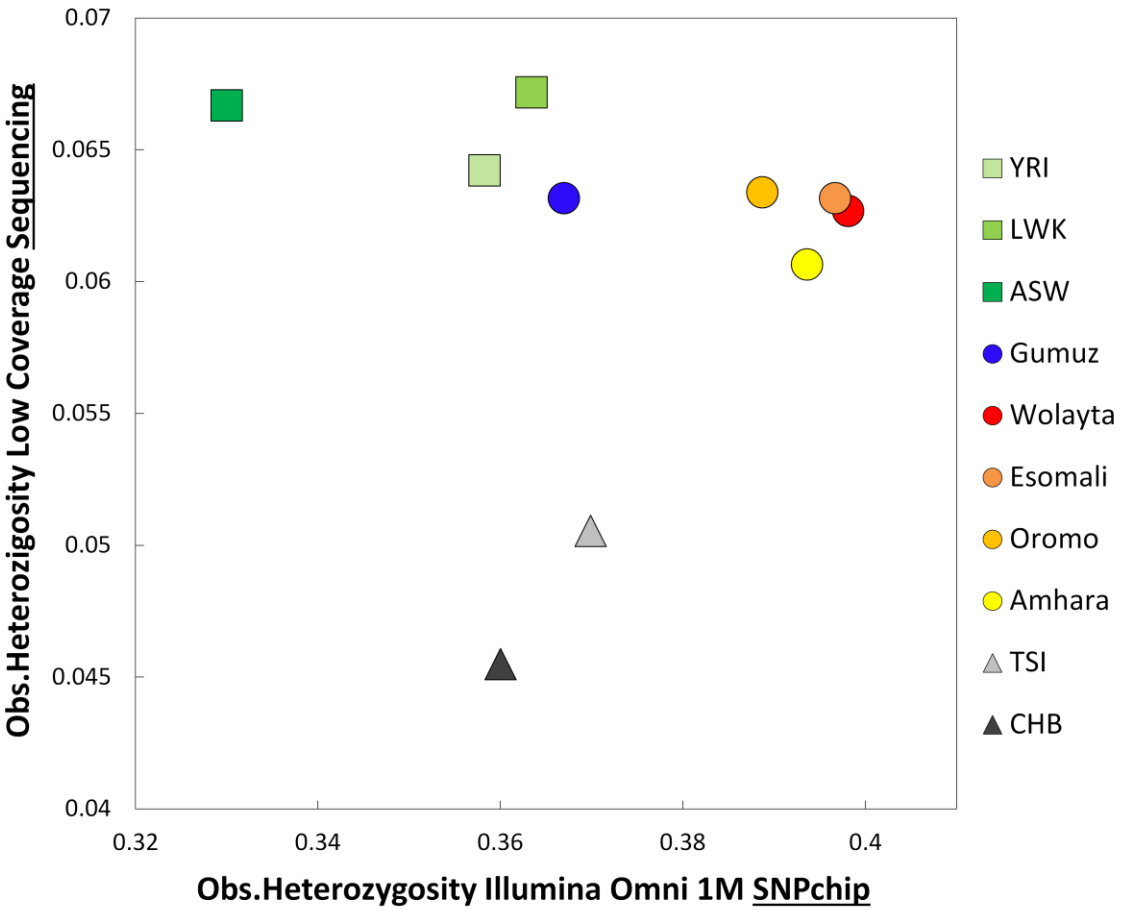

Figure S1: Effect of ascertainment bias in heterozygosity estimation in a set of African and non-African populations. Five Ethiopian populations (Amhara, Oromo, Ethiopian Somali, Wolayta and Gumuz) were chosen from the ones typed in our previous work [1] as the ones maximizing genetic and cultural diversity in the area. The 125 new Ethiopian samples recollected for this study were obtained from 10 ml of peripheral venous blood kindly provided by healthy adult donors who reported having four grandparents belonging to the same ethnic group as themselves. Each donor who signed the informed consent was given an information sheet and briefed about the general aims of the project. Samples were collected in Ethiopia, in the areas or origin of each ethnic group. Each blood sample was extracted within 48 hours from collection using Qiagen FlexiGene DNA Kit. The 100 Egyptian samples were collected in Lebanon from temporary resident Egyptians who trace their ancestry in Egypt for at least the past 3 generations and can therefore be assimilated to a random sample from the broad Egyptian population. Each subject donated 4ml of peripheral venous blood after reading and signing a consent form. DNA was extracted using a standard phenol-chloroform procedure at the Lebanese American University, Lebanon. Ethiopian and Egyptian samples were shipped to the Wellcome Trust Sanger Institute (WTSI), Hinxton, UK for quality checks, genotyping and sequencing. Samples' collection, export and analysis were approved by UK, Ethiopian and Lebanese Research Ethics Committees (UCL REC 0489/002; Ethiopian Ministry of

Science and Technology Approval: 310/538/04 and Lebanese American University IRB SMPZ121307-2 respectively).

The extracted, purified DNA samples were quantified and assessed for integrity, and suitable samples were genotyped on an Illumina Omni 2.5M SNPchip. Whole-genome sequencing was carried out on the same samples using an Illumina HiSeq 2000 platform with a library insert size of 350 bp to an average depth of 8x. These 225 low-coverage samples were pooled together with 837 samples (100 TSI - Toscani in Italy, 91 CEU - Utah residents with Northern and Western European ancestry, 49 IBS - Iberian populations in Spain, 83 LWK - Luhya in Webuye, Kenya, 84 YRI - Yoruba in Ibadan, Nigeria, 49 ASW - African Ancestry in Southwest US, 13 GIH - Gujarati Indian in Houston,TX, 83 CHB - Han Chinese in Bejing, China, 92 CHS - Southern Han Chinese, China, 64 PUR - Puerto Rican in Puerto Rico, 55 CLM - Colombian in Medellin, Colombia, 29 PEL - Peruvian in Lima, Peru) from the 1000 Genomes Project[2] and SNPs called together using the vr-pipe pipeline (https://github.com/VertebrateResequencing/vr-pipe) to avoid batch effects. This calling pipeline, developed as part of the 1000 Genomes Project[2] builds evidence to call heterozygous sites not only from a single sample, but also on the total reads obtained through pooling together all the samples,. Via this approach the vr-pipe pipeline first discovers variable sites, and then assigns a genotype for each site in each sample hence maximizing the calling accuracy obtained from low coverage samples. The called genotypes were phased and imputed using ShapeIT [3] as part of the vr-pipe processing. The quality of calling (96% genotype concordance rate) was estimated by comparing the genotypes from the sequence data with those from the SNPchip. One sample from each of the five Ethiopian groups and three Egyptian samples were also chosen at random after excluding outliers in the PCA and ADMIXTURE profile and further sequenced to a final depth of 30x. The selection of the Ethiopian HC samples was determined by the availability of five distinct Ethiopian populations. Therefore we chose one from each of the five populations. Following this logic, only 1 HC sample would have been chosen to represent the Egyptians. However, due to the larger Egyptian sample size and budget availability we decided to include 3 HC samples for this population. A direct comparison between the genotype calls obtained for these high coverage samples and the same samples down-sampled to 8x coverage, showed a similar genotype concordance rate (95%) as the one obtained from the SNPchip comparison. Furthermore our 95-96% concordance rate is consistent with what previously reported for the same pipeline[2], where 276/287 SNPs (96.2%) were consistently validated by various independent approaches.

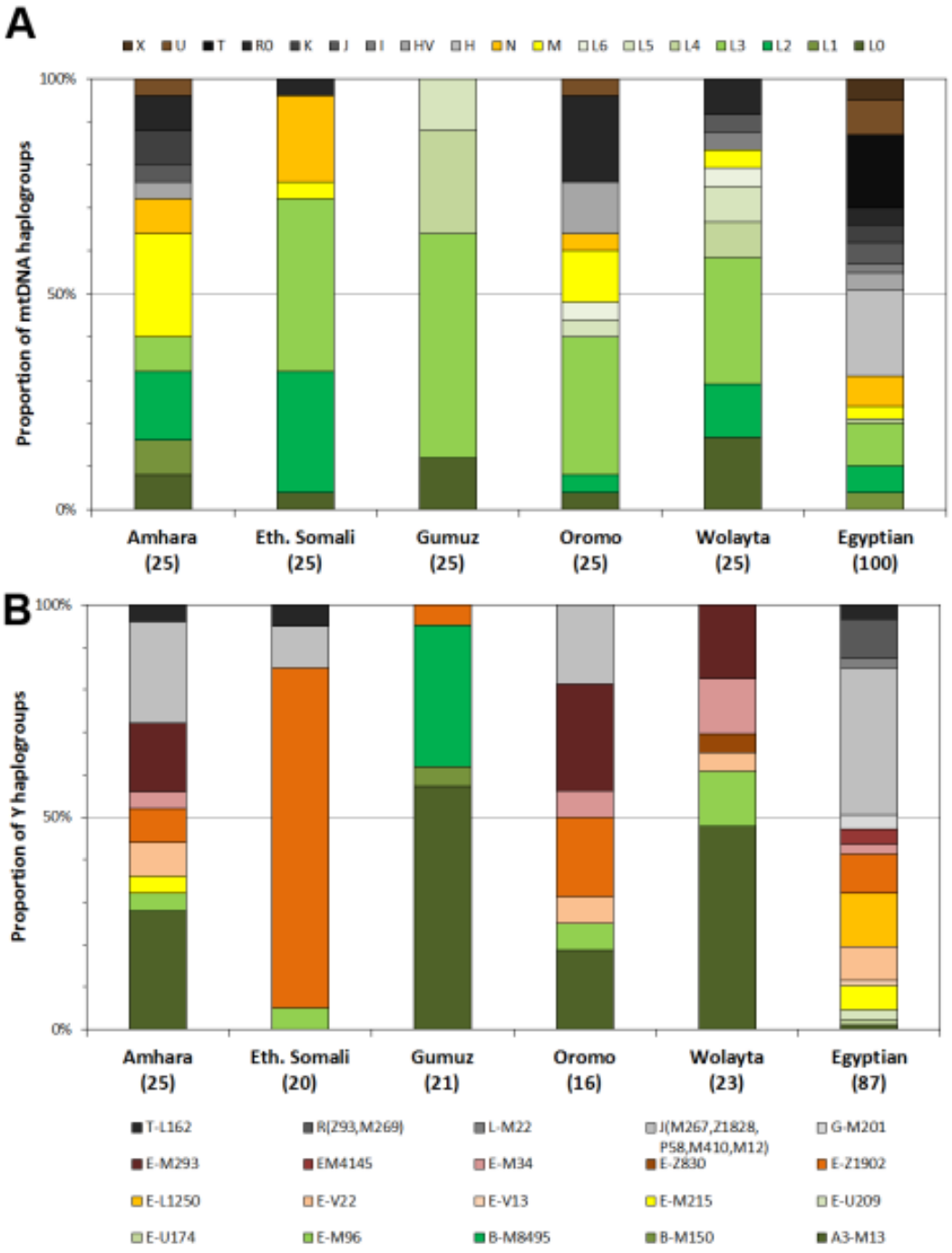

Figure S2. Frequency of African (shades of green) and non African (shades of grey) mtDNA (A) and Ychr (B) haplotypes. Y-chromosome haplogroups were called from the sequenced SNPs using the most up to date nomenclature [4], while the mtDNA haplogroups where inferred from the whole genome fasta sequences [5]. The proportion of non African mtDNA overall fits with the proportion of admixture estimated from the Admixture plot in Figure 1, while the Ychr frequencies show a higher resilience of Y-African lineages to the backflow. Haplogroups whose African/non-African assignation could not be determined are coloured in Yellow and Light Orange. The number of samples available for each population is reported in brackets

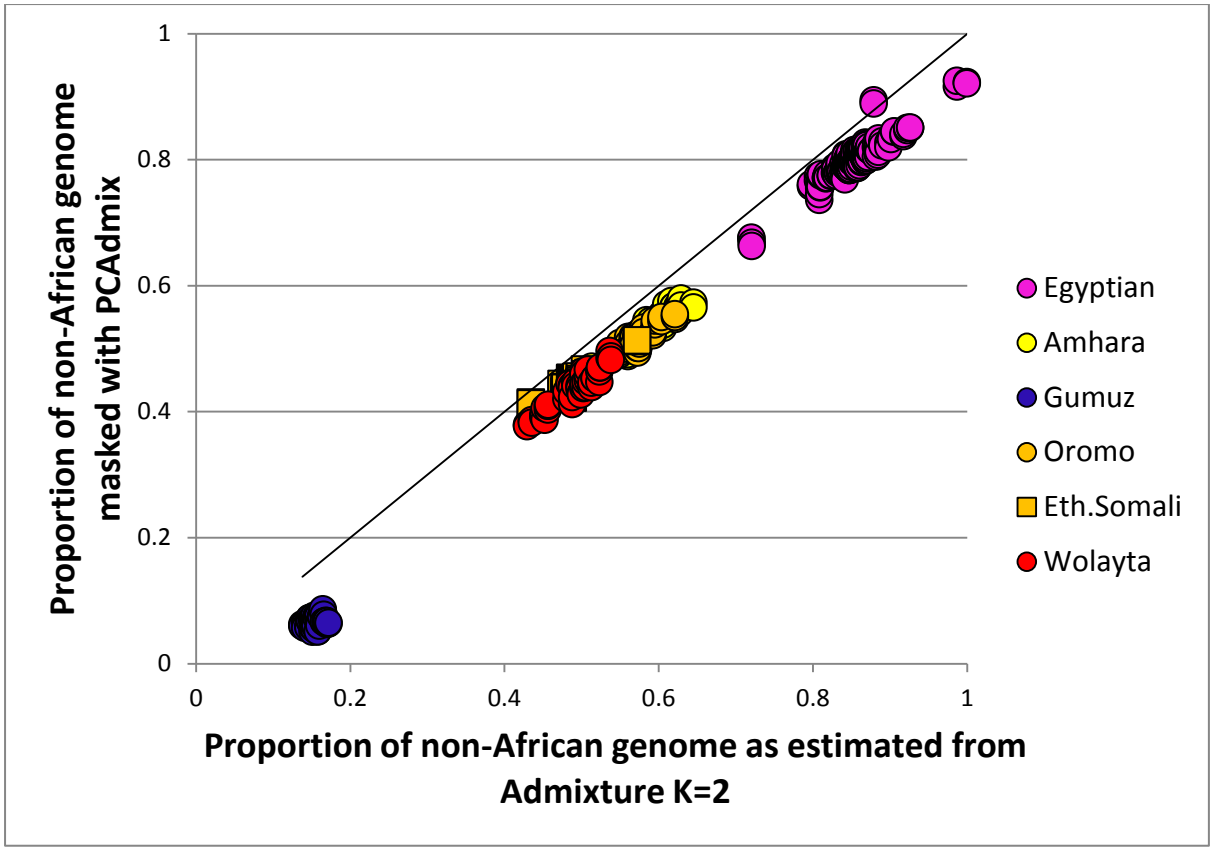

Figure S3 The African and non-African ancestries of the Egyptian, Ethiopian and ASW were deconvoluted using PCAdmix on 20-SNP windows [6]. CEU (which were previously shown[1] to be not significantly different from TSI as surrogate sources for the non African Ethiopian component) for and Gumuz haplotypes were used as reference sources for the non-African and African components, respectively, and the resulting non-African haplotypes masked out from most downstream analyses. The decision to include 20 SNPs per genomic window, as per PCAdmix default parameters, stemmed from the requirement of maximizing the PCAdmix power of resolving ancestry assignation while minimizing the risk of missing out ancestry switches. In the Ethiopian samples, where the Eurasian back flow reportedly took place as long as 85 generations ago, PCAdmix power would be limited by the chunk lengths generated by recombination since the admixture event and with increasing chunk lengths PCAdmix is expected to show decreasing power to correctly recover the fine grained ancestry switch landscape. On the other hand, smaller number of SNPs (e.g. 10) would have decreased the PCAdmix power to discriminate between African and non-African states, as previously shown[7]. The proportion of non African genome masked out in each Ethiopian or Egyptian samples with PCAdmix is compared with the same proportion estimated by Admixture analysis (K=2). The line shows the 1:1 ration, hence highlighting just a slight reduction in the masking efficiency of PCAdmix.

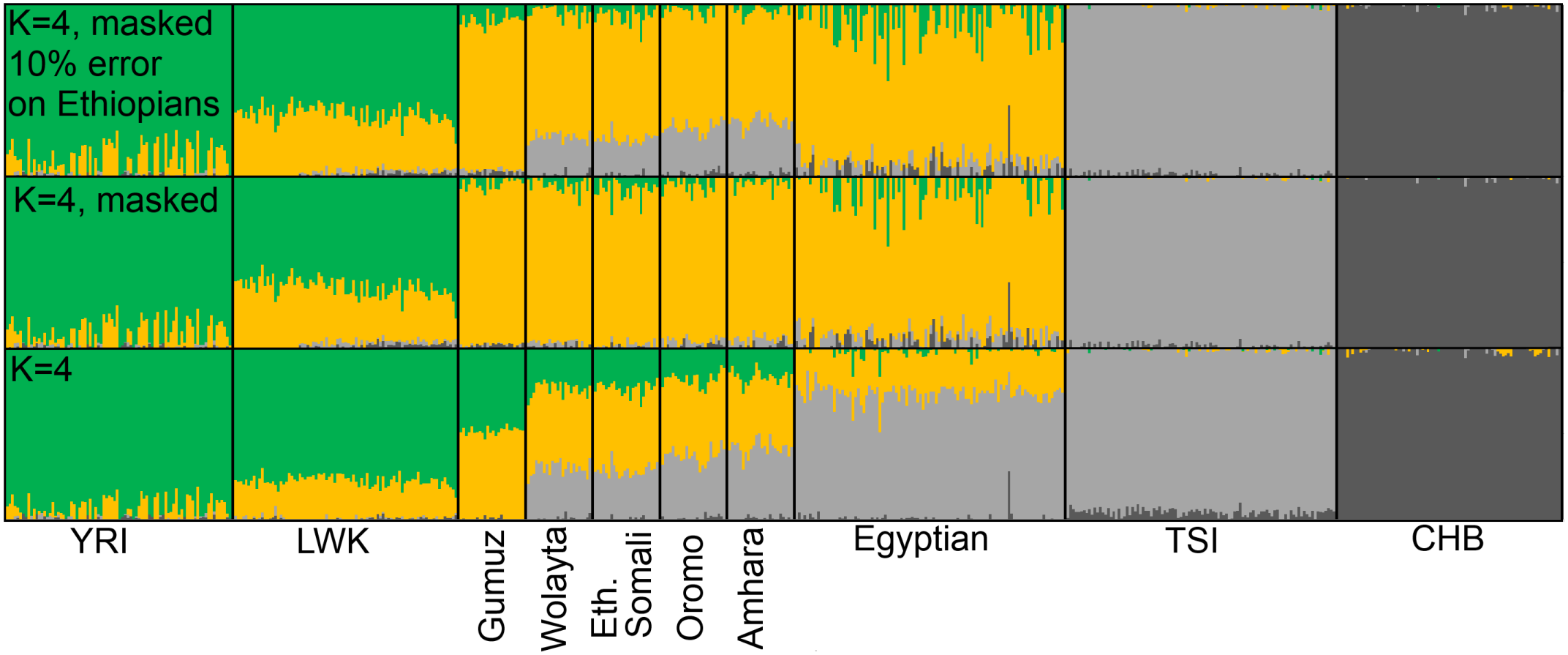

Figure S4. Admixture (K=4) plots showing the change in non-African genomic component (Grey) in Ethiopian and Egyptians before (bottom) and after (middle) ancestry deconvolution and masking. The top admixture plot shows the proportion of non-African genomic component after artificially introducing a 10% masking error in Ethiopians. K4 was chosen to assess the reliability of our masking procedure because K4 is the highest level of K where Eurasian ancestry is detectable in Ethiopians and therefore this was used in some of the downstream analyses as a preferred option over the one suggested by the cross-validation error.

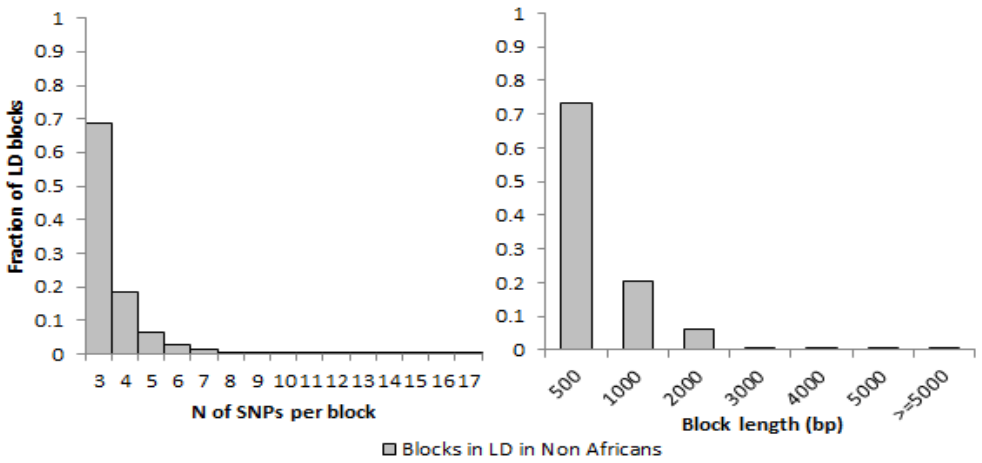

Figure S5: **Length distribution of genome wide Linkage Disequilibrium (LD) blocks inferred from 457 non African samples.** Genome-wide LD blocks defined by three or more SNPs (using Haploview [8] defining blocks assets of SNPs showing all r^2 or D' with the first SNP in the block >=0.8) were identified from 457 non-African phased samples (TSI, CEU, IBS, GIH, CHS, CHB and PEL) spanning a total length of 55 Mbp. The LD blocks were further refined including only sets of SNPs where the r^2 and D' thresholds were met for each pair of markers (i.e. forming a clique), spanning a total length of 7.1 Mbp. The bulk of these blocks were short (≤2000bp) and made up of less than 7 SNPs , consistent with their expected origin pre-dating the time of the OOA migration and surviving the population splits and private drift that occurred thereafter. A list of the OOA haplotypes observed in the Egyptian and Ethiopian samples after masking out their non-African component and in YRI was compiled for each of the LD blocks. To counteract potential bias introduced by the variable numbers of chromosomal regions after ancestry masking at the different genomic locations, the haplotype analysis was constrained for each block to between a minimum of 30 and a maximum of 50 phased chromosomes from each population. Each haplotype was labelled as either shared by two or three populations within Africa, or as unique to a specific African population. The frequency of each haplotype labelled in this way was then computed in the set of non-African populations and the differences in frequencies interpreted as increased or decreased affinity between the non-Africans and Ethiopians, Egyptians or YRI.

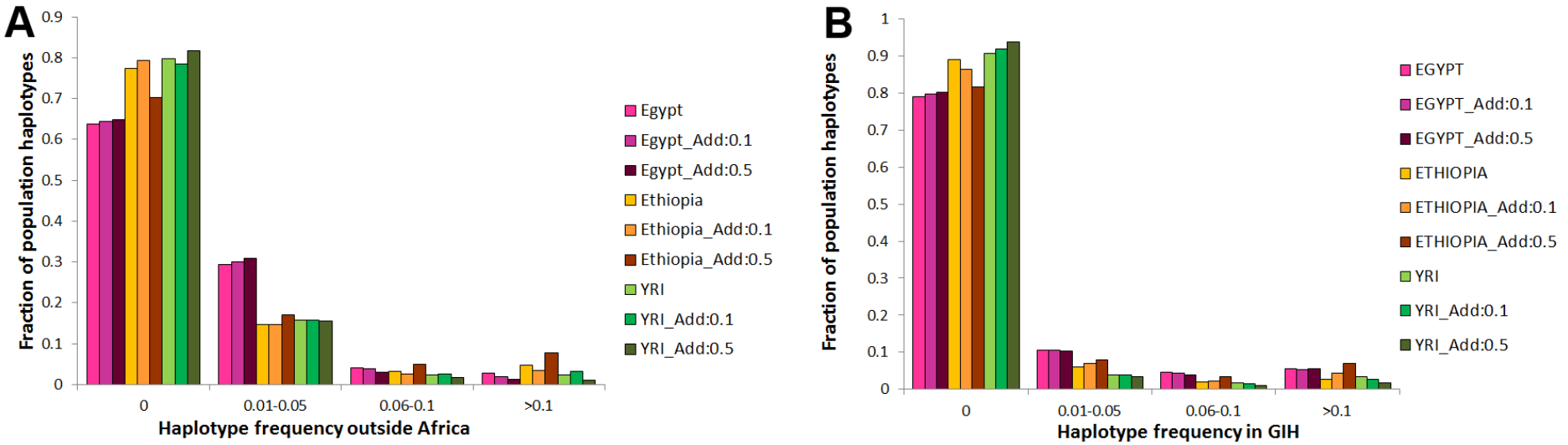

Figure S6 Assessment of technical and biological artefacts in the admixture masking and haplotype sharing analyses. To account for potential masking artifacts in Egyptians, the Ethiopian genomes were masked artificially assigning 10% or 50% of the inferred CEU sites to the Ethiopian African component (A). The same process was also carried out using GIH (B) instead of TSI+CHB (A) as the non African population. In both cases the results showed Egyptians as the African population closest to the non Africans.

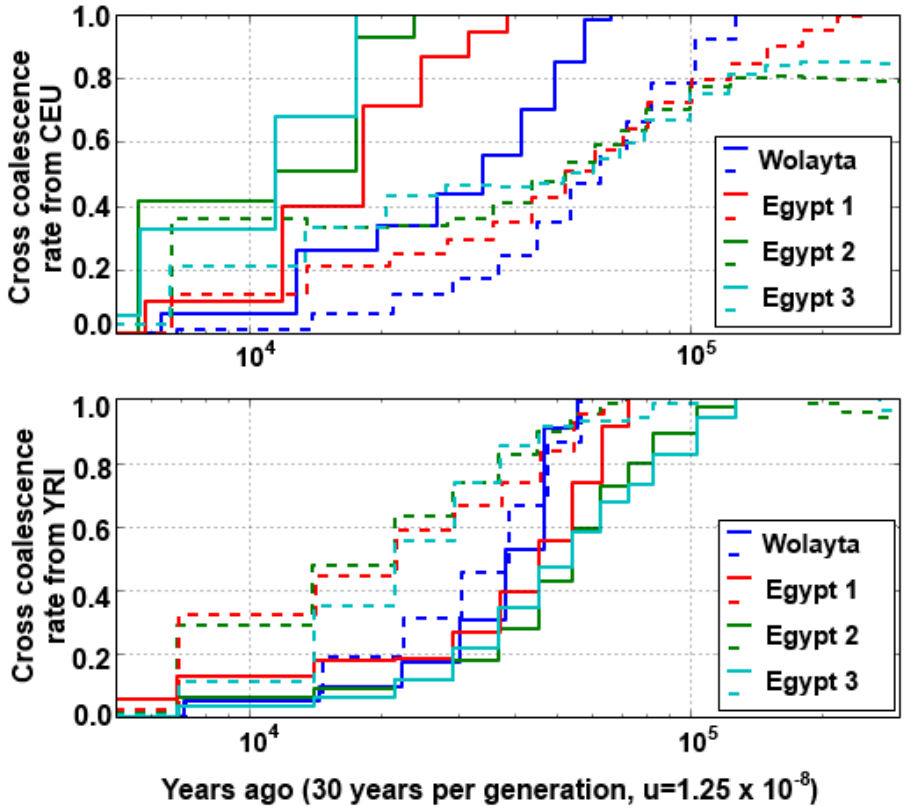

Figure S7. MSMC-plots of inferred genetic split times (coloured lines) between 1 Ethiopian (Wolayta) and 3 Egyptians from CEU (top panel) and YRI (bottom panel). The same splits were calculated also after masking out their non-African component (dashed lines). The split between two genomes is defined as the time when the cross-coalescence rate drops to 50%. The five Ethiopian and three Egyptian were randomly chosen and sequenced at high-coverage (30x) . These genomes were processed using the MSMC [9] pipeline (https://github.com/stschiff/msmc) and the called genotypes combined with those from a set of reference high-coverage genomes [2; 10] to estimate the relative genetic split times (using 30 years per generation and a mutation rate of 1.25 10^{-8} per bp per generation). MSMC was run on the same Egyptian and Ethiopian samples before and after removing the non-African genomic components

Biological_Consequence	YRI	LWK	ASW	Gumuz	Wolayta	Amhara	Oromo	Eth.Somali	Egyptian	CEU	TSI	CHB	CHS
mature_miRNA_variant	133	131	140	93	118	117	119	111	97	78	81	73	63
TF_binding_site_variant	2718	2923	2990	2491	2743	2619	2662	2459	2209	1693	1770	1519	1505
regulatory_region_variant	281070	296156	302883	255513	277956	265834	272949	255832	224657	175589	180305	158702	156064
5_prime_UTR_variant	18285	18772	19477	17258	18874	17906	18712	17565	15305	11643	12067	10138	9752
splice_acceptor_variant	221	235	249	200	225	207	226	193	184	141	143	135	138
splice_donor_variant	370	381	388	329	358	348	372	347	291	255	252	219	219
intron_variant	5669433	6021481	6150595	5098882	5580291	5322616	5499555	5150386	4513470	3521768	3587977	3171249	3139034
missense_variant	36868	39149	40563	34317	38315	37109	38152	35218	29617	23894	24700	21128	20321
synonymous_variant	40738	42706	43708	37057	40486	38912	39994	37382	32410	25073	26118	22400	21886
non_coding_exon_variant	69913	73772	75024	64143	69422	66360	68411	64488	56442	43830	44933	38854	38417
stop_gained	325	375	395	311	344	348	338	316	246	224	230	209	197
stop_lost	47	43	46	41	42	48	40	44	34	29	29	21	28
3_prime_UTR_variant	87034	92257	94393	77570	85923	81938	84772	79363	68462	53931	54831	47951	47390
intergenic_variant	4138231	4386001	4465844	3708584	4041408	3860912	3982312	3746811	3322590	2589775	2639431	2350418	2330787
Tot Variants	10345386	10974382	11196695	9296789	10156505	9695274	10008614	9390515	8266014	6447923	6572867	5823016	5765801

Table S1 Variant annotation based on 22 Low Coverage samples from each population. Cells are color-coded on a scale going from red (high values) to green (low values).

Recipient pop	Source pop1	Source pop2	Z score of f3 test for admixture	Years since admixture (+/-Error)
Gumuz	Yoruba (YRI)	Tuscan (TSI)	51.0433	NA
Egyptian	Yoruba (YRI)	Tuscan (TSI)	-97.3065***	772.5 +/- 25.8
Egyptian	Gumuz	Tuscan (TSI)	-87.8906***	778.8 +/- 27.9
Oromo	Gumuz	Tuscan (TSI)	-103.402***	1782.9 +/- 129.9
Wolayta	Gumuz	Tuscan (TSI)	-93.7798***	1788.6 +/- 111.3
Amhara	Gumuz	Tuscan (TSI)	-103.547***	2447.7 +/- 117.9
Somali	Gumuz	Tuscan (TSI)	-52.8217***	3641.4 +/- 191.7

Table S2 f3[11] and ALDER[12] admixture results. Trios of populations (Recipient Pop;YRI/Gumuz,TSI) showing significant signs of admixture (Z score ≤-2) are labelled with "***" and the midpoint of the admixture event dated with Alder. When excluding the Oromo from this calculation due to the reported double admixture event[13] which is likely to bias toward the present point estimate, the average Eurasian backflow in Ethiopia can be dated at 2625 years ago.

	YRI	ASW	Gumuz	Wolayta	Somali	Amhara	Oromo	Egypt	CEU	CHB
YRI		0.006	0.020	0.020	0.020	0.019	0.020	0.018	0.072	0.041
ASW	0.007		0.021	0.021	0.021	0.019	0.021	0.019	0.062	0.037
Gumuz	0.018	0.015		0.023	0.023	0.017	0.024	0.024	0.108	0.061
Wolayta	0.019	0.015	0.014		0.020	0.022	0.021	0.023	0.109	0.061
Somali	0.019	0.015	0.014	0.010		0.022	0.021	0.022	0.111	0.062
Amhara	0.018	0.014	0.011	0.013	0.012		0.023	0.023	0.099	0.055
Oromo	0.020	0.015	0.015	0.011	0.011	0.013		0.023	0.107	0.061
Egypt	0.025	0.017	0.018	0.015	0.013	0.016	0.013		0.111	0.060
CEU	0.036	0.026	0.031	0.026	0.024	0.027	0.024	0.008		0.030
CHB	0.041	0.033	0.041	0.037	0.036	0.037	0.036	0.024	0.060	

Table S3 Pairwise F_{ST} before (lower triangle) and after (upper triangle) masking out the non-African portion of the Egyptian, Ethiopian and Afro-American (ASW) samples analysed in this study. Cells are color-coded on a scale going from red (high values) to green (low values).

True ancestral	Input reference	Proportion (SD) of European admixture simulated	Mean proportion actually in simulated samples	Time of admixture (generations)	Correlation between simulated proportional ancestry and PCAdmix ancestry	Accuracy (sites matching true assignment)	**Non-African ancestry misclassified as African (% of all sites)- False negatives**	African ancestry misclassified as non-African (% of all sites)- False positives	Proportion of CEU ancestry misclassified (%)	Proportion of African ancestry misclassified
CEU, GUMUZ	CEU, GUMUZ	0.50(0.03)	0.51(0.11)	85	0.94	82.6%	**4.0%**	13.3%	7.6%	27.2%
CEU, GUMUZ	CEU, GUMUZ	0.80 (0.04)	0.79(0.13)	25	0.87	88.0%	**6.6%**	5.0%	8.5%	24.5%
TSI, GUMUZ	CEU, GUMUZ	0.80(0.04)	0.77 (0.14)	25	0.91	87.3%	**6.7%**	5.9%	5.3%	25.9%
CEU, GUMUZ	CEU, YRI	0.50(0.03)	0.51(0.11)	85	0.92	79.6%	**5.4%**	15.0%	10.4%	30.7%
TSI, GUMUZ	CEU, GUMUZ	0.50(0.03)	0.52(0.09)	85	0.89	84.4%	**2.5%**	13.0%	4.8%	27.2%

Table S4 Estimation of masking error with PCAdmix. To assess the potential effect of artefacts introduced by the masking process, we generated ten admixed genomes by mixing CEU or TSI and GUMUZ or YRI haplotypes in various proportions and introducing ancestry switches to recreate the conditions observed in our Ethiopian and Egyptian data. We then tried to mask these artificial genomes with PCAdmix with the same settings described above and using either CEU and YRI or CEU and GUMUZ as putative source populations, making sure none of the source individuals were also used to generate the artificial genomes. In the 'Egyptian' case (20% African component, 25 generations since the admixture) we observe the highest proportion (6.7%) of Non-African ancestry misclassified as African (column 8, in bold). This misclassification in Egyptians could have introduced a bias in the estimation of similarity between them and OOA, however since a similar bias is also observed for the Ethiopian artificial population (50% African component, 85 generations since the admixture) the two effects are likely to cancel each other out. To counterbalance potential over-estimation of the Egyptian-OOA similarity, we re-run the haplotype sharing analyses artificially introducing a 10% (Figure S4) or 50% misclassification error in Ethiopians. By this way we increased the proportion of CEU sites misclassified as African in Ethiopians, therefore counterbalancing the potential bias observed in the Egyptian case. While the introduction of artificial 10% and 50% CEU mis-assignation produced a shift toward higher affinity between Ethiopians and the OOA populations, the qualitative finding of Egyptian being the African population closest to the assessed non-Africans did not change (Figure S6a).Furthermore we also re-run the haplotype sharing analysis using GIH (a South Asian population placed along the putative southern route out of Africa). The results for GIH replicated what already found for the CHS + TSI populations (Figure S6b) and are provided only as supplementary results due to the small sample size (N=13) available from the Phase 1 of the 1000 Genomes Project [2].

Supplementary References

1. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., et al. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am J Hum Genet 91, 83-96.
2. The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
3. Delaneau, O., Marchini, J., and Zagury, J.F. (2012). A linear complexity phasing method for thousands of genomes. Nat Methods 9, 179-181.
4. van Oven, M., Van Geystelen, A., Kayser, M., Decorte, R., and Larmuseau, M.H. (2014). Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat 35, 187-191.
5. 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.
6. Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J.G., and Bustamante, C.D. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol 84, 343-364.
7. 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.
8. Barrett, J.C. (2009). Haploview: Visualization and analysis of SNP genotype data. Cold Spring Harb Protoc 2009, pdb ip71.
9. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat Genet.
10. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327, 78-81.
11. Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., and Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet 7, e1001373.
12. Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193, 1233-1254.
13. Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A 111, 2632-2637.