Genetic Study of Population Mixture and Its Role in Human History

Genetic study of population mixture and its role in human history

A dissertation presented

by

Priya Moorjani

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics

Harvard University

Cambridge, Massachusetts

May 2013

Dissertation Advisor: Professor David Reich Priya Moorjani

Genetic study of population mixture and its role in human history

Abstract

Mixture between populations is an evolutionary process that shapes genetic variation. Intermixing between groups of distinct ancestries creates mosaics of chromosomal segments inherited from multiple ancestral populations. Studying populations of mixed ancestry (admixed populations) is of special interest in population genetics as it not only provides insights into the history of admixed groups but also affords an opportunity to reconstruct the history of the ancestral populations, some of whom may no longer exist in unmixed form. Furthermore, it improves our understanding of the impact of population migrations and helps us discover links between genetic and phenotypic variation in structured populations.

The majority of research on admixed populations has focused on African Americans and Latinos where the mixture is recent, having occurred within the past 500 years. In this dissertation, I describe several studies that I have led that expand the scope of admixed studies to West Eurasians and South Asians where the mixture is older, and data from ancestral groups is mostly unavailable. First, I introduce a novel method that studies admixture linkage disequilibrium (LD) to infer the time of mixture. I analyze genomewide data from 40 West Eurasian populations and show that all Southern European, Levantine and Jewish groups have inherited sub-Saharan African ancestry in the past 100 generations, likely reflecting events during the Roman Empire and subsequent Arab migrations.

Next, I apply a range of methods to study the history of Siddi groups that harbor African, Indian and Portuguese ancestry, and to infer the history of Roma gypsies from Europe.

Finally, I develop a novel approach that combines the insights of frequency and LD-based statistics to infer the underlying model of mixture. I apply this method to 73 South Asian groups and infer that major mixture occurred ~2,000-4,000 years ago. In a subset of populations, all the mixture occurred during this period, a time of major change in India marked by the de-urbanization of the Indus valley civilization and recolonization of the Gangetic plateau.

Inferences from our analyses provide novel insights into the history of these populations as well as about the broad impact of human migrations.

Table of Contents

List of Figures and Tables

Chapter 5: Genetic evidence for recent population mixture in India

Appendix A: Supplementary Material for Chapter 2

Appendix B: Supplementary Material for Chapter 3

Appendix C: Supplementary Material for Chapter 4

Appendix D: Supplementary Material for Chapter 5

Acknowledgements

This work would not have been possible without the help and support of many people. First and foremost, I would like to thank my advisors, David Reich and Nick Patterson, for their great mentorship and tremendous support, both scientifically and personally. Their creativity and enthusiasm for research has been a great source of inspiration for me. Working with David has taught me to think critically as a scientist, identify interesting scientific problems, and communicate my research ideas clearly and precisely to a wide audience. I have enjoyed learning about a wide range of topics from statistics to politics from Nick. I thank you both for guiding me on the path to becoming an independent researcher.

I would like to thank my collaborators, in particular Harry Ostrer, Béla Melegh, Lalji Singh and Thangaraj Kumaraswamy, without whom this research would not have been possible. I would also like to thank my thesis committee members, George Church, Shamil Sunyaev, and Alkes Price for your valuable advice and critiques. I am grateful for the discussions during the thesis meetings, which always led to new ideas and directions to follow.

I would like to thank all the members of the Reich lab for their help and support over the past four years. I will always fondly remember our stimulating lunch discussions and teatime, which made coming to the lab everyday fun and exciting. I would like to thank my office mates, Suzanne Nordenfelt and Kasia Bryc, for always lending an ear and providing great advice about all life matters. Special thanks to Elizabeth Fels for keeping the lab organized and magically fixing every problem (sometimes even before they occur). Thank you to Sriram Sankararaman, Kasia Bryc, and Joseph Pickrell for their constructive comments on the dissertation.

My warmest thanks to my family – my sister, Samira Moorjani, my brother-in-law, Jatin Sonavane and uncle, Kishin Moorjani – for their support during my graduate career. I would also

like to thank my friends, in particular Colm O'Dushlaine, Tina Liu and Vedangi Sample, who have made Boston my home away from home. Vedangi, I am really happy that we reconnected in Baltimore; I have tremendously cherished your friendship in the past 8 years. Special thanks to my boyfriend, Denis Titov, for his unrelenting patience in listening to my never-ending stories about work, for his endless support and encouragement, and for sharing all my crazy interests - be it traveling or eating spicy hell bones. I look forward to many more adventures with you.

And finally, I would like to thank my mother, Neena Moorjani – for her unquestioning love and support throughout my life, for her many sacrifices, and for always believing in me (even when I doubted myself) – I dedicate this thesis to you.

To my mother

for her unconditional love and support

Chapter 1

Introduction*

**Some of the material was originally published as:* Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. (2013) Ancient admixture in human history. *Genetics*. 2012 Nov;192(3):1065-93. doi: 10.1534/genetics.112.145037. Epub 2012 Sep 7. It has been modified to fit the style of the thesis.

The fields of archaeology, anthropology, linguistics, and the study of written texts have, until recently, provided the main sources of information about inferences of human population history. In the past few decades, however, population genetics – the study of inherited genetic variation among populations – has established itself as a powerful new tool in this armamentarium. The main kinds of genetic variation observed in humans are – changes in single nucleotides (single nucleotide polymorphism (SNP)), changes in the number of copies of the sequence (copy number variation (CNV)), and insertions or deletions of the genomic sequence (InDels). These are primarily caused by evolutionary processes such as mutation, recombination, and gene flow. Changes in the genome can affect the phenotype (such as skin color, disease traits, height etc) in an individual. The first few population genetics studies utilized this insight and indirectly investigated genetic variation by examining the distribution of phenotypes such as blood groups and observed that different human populations have different proportions of blood groups[1]. However, with the advent of electrophoresis, it became possible to directly characterize genetic markers into different forms or alleles[2; 3]. By studying the distribution of these alleles across different populations, for the first time, one could infer the history of populations directly using genetic data[1].

These studies were soon followed by sequencing of haploid, uniparentally-inherited markers, e.g. Y-chromosome and mitochondria, that can be used to trace an individual's ancestral (paternal or maternal) lineages back in time[4-6]. These studies not only provided new insights about the settlement of the world, for example by showing that the main history of human populations is a story of migration and dispersal out of Africa ~45,000-60,000 years ago[7-9], but also provided a novel way of investigating the evolutionary history of our species.[8; 10; 11] A limitation, however, was that these inferences were often based on a handful of markers, and so

had limited resolution. For example, recent studies examining hundreds of thousands of markers on autosomal DNA (chromosome 1-22) have shown that most non-Africans have ~1-4% Neandertal ancestry[12]; a signal previously missed by studies of mitochondrial markers[13].

Technological and scientific advances that followed made large-scale, genome-wide investigations feasible and cost-effective. This enabled studies of fine-scale population structure[14; 15], estimation of basic parameters like mutation rate[16; 17] and recombination rate[18; 19], impact of human adaptation to new environments[20], as well as identification of disease risk variants[21]. Another important observation from these studies was that populations did not simply separate once and occupy different geographic locations; they had many secondary contacts (after the initial migration out of Africa) and there was often mixture between individuals of different groups, such that most populations around the globe today trace their ancestry to more than one distinct ancestral group[1; 22]. These populations of mixed ancestry are referred to as admixed populations. For example, African Americans have on average about 80% African ancestry and 20% European ancestry because of mixture that occurred in the past 500 years[23], and most Latinos harbor Native American, European and African ancestry[24]. There are only a handful of human populations today that do not show any evidence of genetic admixture, and in these populations, it is possible that the only reason there is no evidence of genetic admixture is simply because it has not been detected so far.

1.1 Characterization of genomes of admixed individuals

The genome of an individual with mixed ancestry (admixed individual) represents a mixture of alleles inherited from multiple ancestral (or parental) populations. Consider two

ancestral populations (shown in blue and orange in Figure 1.1) that diverged from each other a long time ago, such that they have very distinct ancestry along their chromosomes. When these populations mix with each other, the first generation offspring inherits one blue and one orange chromosome. Recombination or crossovers in admixed individuals breaks down these contiguous ancestry blocks, and leads to the creation of mosaic of chromosomal segments of distinct (blue or orange) ancestry. The time and amount of gene flow between ancestral populations influences the genetic variation (through the distribution of blue and orange ancestry blocks) in the admixed population[25; 26]. In admixed individuals, markers show extended allelic correlation or linkage disequilibrium (LD) relative to the ancestral populations and rate of decay of the LD is related to the proportion of mixture, recombination rate and the time since admixture[27]. These parameters (the time and amount of admixture) are not directly observed but can be inferred from genetic data. Thus, studying the patterns of genetic variation and extended LD in admixed populations can be useful for learning about human history.

1.2 Methods for analyzing population mixture

Over the past few decades, many methods have been developed for the identification of the source of the admixing populations, estimation of the proportion of ancestry derived from each ancestral group, and inferring signatures of admixture linkage disequilibrium and the time of the admixture event(s). These methods broadly fall into two categories: (a) Descriptive methods, and (b) Inference methods.

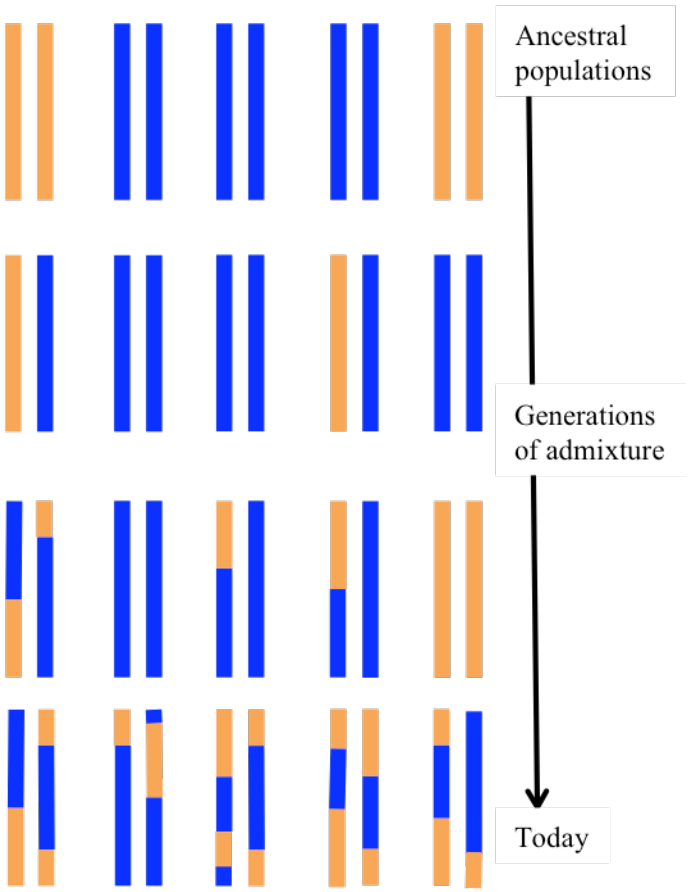


Figure 1.1 Schematic of genomic ancestry resulting from admixture between two ancestral groups (shown in blue and orange). Each chromosome is represented by a vertical line. Gene flow between the ancestral groups creates mosaic chromosomes containing blocks of ancestry (either blue or orange) that vary in length depending on recombination rate and the time since mixture. This figure was adapted from references [22] and [25].

Descriptive methods

Clustering methods such as Principal Component analysis (PCA)[28; 29], STRUCTURE[30] and ADMIXTURE[31] have become indispensible tools for investigating population structure in genomic data. These methods are based on identifying the most significant components of variation in the data and highlighting the patterns of similarities and differences between

samples. Using STRUCTURE, Rosenberg et al. (2002) showed there is continental and regional level clustering in genetic data, and by applying PCA[32], Novembre et al. (2010) showed that one can re-create the geographical map of Europe by studying the genetic variation in European populations[33]. While these methods are powerful for detecting population substructure, they do not provide any formal tests (the patterns in data can be generated by multiple population histories). For instance, Novembre and Stephens (2010) showed that isolation-by-distance – in which nearby populations exchange more genes with each other than with populations located farther away – could generate gradients of genetic variation that look similar to long-distance historical migrations[34]. STRUCTURE/ADMIXTURE results are also difficult to interpret historically. These methods either work without explicitly fitting a historical model or by fitting a model that is (usually) unrealistic as it assumes that all the populations have radiated from a single ancestral group.

Inference methods

Inference methods broadly fall into two classes: local ancestry-based and global ancestry-based methods, based on the kind of information – allele frequency, LD or haplotype – they use for making inferences. Local ancestry-based methods such as LAMP[35], HAPMIX[36], FineSTRUCTURE[37] and PCADMIX[38] deconvolve ancestry at each locus in the genome and provide individual-level information about ancestry. Recent methods by Pool and Nielson (2011), Pugach et al. (2011) and Gravel et al. (2012) study the ancestry tract lengths distributions to estimate a date of mixture[39-41]. These methods rely on accurate local ancestry information, which becomes harder when the reference populations are highly divergent from the true mixing

populations or the ancestry tracts are short (as in the case for ancient admixtures). These methods can provide valuable insights into the recent history of populations, but they have reduced power to study older events such as the ones I investigate in this dissertation (~30-300 generations).

Another approach for studying individual level variation is to study the distribution of Identity-by-descent (IBD) sharing among individuals. Methods such as GERMLINE[42], PLINK[43] and BEAGLE FastIBD[44] compare pairs of individuals to find shared tracts of chromosomal segments that are inherited identical by descent due to recent common ancestry. Inferred IBD segments provide information about the relationships between individuals – individuals that share many long IBD segments, most likely have common ancestry within the past few generations. This insight was used by Atzmon et al. (2011) to show that diverse Jewish groups trace some of their ancestry to a recent common ancestor[45]. My study on the Roma gypsies examines the distribution of IBD sharing between the Roma and non-Roma populations in India to infer that the Roma are most closely related to Northwest Indian populations (Chapter 4). A limitation of IBD methods is that they require large datasets (with many samples and high density of SNPs) for accurate detection of IBD segments. In addition, current methods can only reliably detect IBD segments that are > 2cM, which limits the utility of these methods to studies of recent events (that occurred within the past 50 generations)[46].

Global ancestry-based methods fit demographic models or phylogenetic trees to population level data. One approach is to explicitly fit a model of demography to genetic data that captures changes in population sizes, times of population mixtures and splits, and other relevant parameters. The values of these parameters can be estimated from the data, often involving simulations. Approximate Bayesian clustering methods[47-49], Isolation Migration models[50-52] and more recently methods by Gutenkunst et al. (2009) and Gronau et al. (2011) are

based on this approach[53; 54]. While these methods provide valuable information about a range of parameters, they can be computationally intractable for large datasets containing many populations.

Another class of global ancestry-based methods uses allele frequency correlations across populations to build phylogenetic trees of population relationships. These methods are inspired by the ideas from Cavalli-Sforza and Edwards (1967), who fit phylogenetic trees of population relationships to the F_{st} values measuring allele frequency differentiation between pairs of populations[55]. These methods do not explicitly model each demographic parameter (the details get absorbed into the branch lengths of the phylogenetic trees), and are thus feasible for examining many populations simultaneously. A limitation of this approach, however, is that mixture between populations violates the tree assumption. Later studies by Thompson (1975); Lathrop (1982); Waddell and Penny (1996); Beerli and Felsenstein (2001); Pickrell and Pritchard (2012), and Patterson et al. (2012) use phylogenetic graphs (phylogenetic trees that allow for admixture) to model allele frequencies observed in multiple populations and provide formal tests to examine the violation of the tree model[56-61]. However, to be computationally tractable, these methods ignore allelic correlation across markers.

Chakraborty and Weiss (1988) introduced an approach that uses associations between pairs of markers to infer demographic parameters related to admixture, such as the proportion and time of admixture[62]. This approach is similar in spirit to estimating dates of admixture based on local ancestry; however, it does not require explicit identification of each ancestry track and is mathematically more tractable. My recent work further develops these ideas and shows that the inferences based on admixture LD can provide unbiased estimates of the dates of mixture up to 500 generations in simulations (Chapter 2, 4)[63; 64]. In addition, recent work by Loh et al. (2013)

provides a framework for using the amplitude of the decay of admixture LD to learn about various demographic parameters such as proportion of mixture from the ancestral population[64], source of the ancestral populations (unpublished data), and underlying model of admixture (Chapter 5).

1.3 Applications of studying admixed populations

Studies of admixed populations can provide invaluable insights about the impact of population migrations. Recent studies have shown that there were three streams of Asian gene flow into Native Americans[65] and that aboriginal Australians are descendants of an early migration out of Africa that possibly dates back to 62,000 to 75,000 years before present[66]. By tracing the source of the ancestry in an admixed population, one can not only understand the history of the admixed population, but also infer the history of the ancestral groups, some of which may no longer be extant. For example, one can rebuild the genomes of Native American populations that have contributed ancestry to the genomes of present day Latinos and Mexicans, but are no longer extant in unadmixed form[24]. In addition, by studying the extent of linkage disequilibrium in admixed groups, one can estimate the time of the admixture event. Sankararaman et al. (2012) apply LD-based methods similar to *ROLLOFF* (Chapter 2) and estimate that the Neandertal admixture occurred in Europeans between 37,000–86,000 years before present[67]. This is roughly consistent with the archaeological evidence for the presence of Neandertal material in West Eurasia during this period[67].

Studying admixed genomes can also be useful in identifying genes related to diseases. Rife (1954) first proposed the use of admixture LD to characterize the genetic basis of traits in

admixed groups[68]. Later work by Chakraborty and Weiss (1988), Stephens (1994), McKeigue (1997), Patterson et al. (2004) and Montana and Pritchard (2004) laid out the foundation of admixture mapping that examines the variation of ancestry along the genome and its correlation with disease risk[25; 62; 69-72]. This approach takes advantage of the fact that near a disease-causing locus there will be enhanced ancestry of the population that has greater risk of causing the disease. Admixture mapping has been very successful in identifying risk variants related to a range of diseases in African Americans and Latinos[25; 73; 74]. However, this technology is currently limited to recently admixed populations only (where the mixture occurred within the past 500 years). New genetic variants introduced through admixture, from one population into another, can have an impact on phenotypic traits. For example, a recent study in Latinos has shown that some of the risk variants associated to diabetes trace their origin to Neandertal admixture in this population (unpublished work)[75]. Also a study of inheritable cardiomyopathies in India has discovered a strong association of this disease to a 25 base pairs (bp) deletion that is common in South Asia, present in Southeast Asia but absent everywhere else in the world[76]. The deletion is likely related to Ancestral South Indian ancestry in Indians. Thus, studying ancestry and its association to disease risk in admixed populations can be very informative for identifying disease variants.

1.4 Summary of Thesis

The majority of research on admixed populations has focused on recently admixed groups such as Africans Americans and Latinos (where the mixture between the ancestral groups has occurred within the past 20 generations). Limited work has been done for making these

methods accessible to other admixed groups, where the admixture is older and/or data from the ancestral population is not available. In this dissertation, I develop and apply novel methods to expand the scope of admixed studies to West Eurasians and South Asians.

First, I present a novel method called *ROLLOFF* that uses admixture LD to infer the date of the admixture event. I perform extensive simulations to show that *ROLLOFF* produces unbiased estimates of the dates of admixture as well as accurate standard errors up to 500 generations and is robust to substantially inaccurate ancestral populations as well as errors in the genetic map at these time scales. I examine allele frequency correlations across populations (f-statistics) to understand population relationships of the admixed groups with other worldwide populations. Finally, I develop a novel approach that combines the inferences from both frequency-based and LD-based methods to infer the underlying model of admixture. This approach compares the amount of LD observed in the target population to the amount expected under a model of single pulse of mixture. This enables me to disentangle models of single vs multiple gene exchanges, while leveraging information from multiple sources of data.

I apply these methods to study the history of sub-Saharan African ancestry in Europeans, Levantines and Jews (Chapter 2); to infer the history of Siddis living in India, who have African, Indian and Portuguese ancestry (Chapter 3); to understand the history of European Roma gypsy populations (Chapter 4); and finally to reconstruct the history of South Asian populations who descend from a mixture of two highly divergent ancestral populations: Ancestral North Indians (ANI) related to West Eurasians, and Ancestral South Indians (ASI) not closely related to groups outside the subcontinent (Chapter 5).

In all four cases, I record strong evidence of admixture using f-statistics and perform analyses to characterize the admixture by inferring the contribution of ancestry from each ancestral population, identifying the populations that are most closely related to the source of the admixture, and estimating the timing of the admixture event. Inferences from these analyses can be used to improve the understanding of population relationships and expand techniques such as admixture mapping to South Asians and West Eurasians.

1.5 References

1. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The history and geography of human genes.(Princeton university press).

2. Pauling, L., Itano, H.A., Singer, S., and Wells, I.C. (2004). Sickle Cell Anemia, a Molecular Disease1. Landmarks in Medical Genetics: Classic Papers with Commentaries, 200.

3. Ingram, V.M. (1957). Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature 180, 326-328.

4. Denaro, M., Blanc, H., Johnson, M., Chen, K., Wilmsen, E., Cavalli-Sforza, L., and Wallace, D. (1981). Ethnic variation in Hpa 1 endonuclease cleavage patterns of human mitochondrial DNA. Proceedings of the National Academy of Sciences 78, 5768.

5. Johnson, M.J., Wallace, D.C., Ferris, S.D., Rattazzi, M.C., and Cavalli-Sforza, L.L. (1983). Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. Journal of Molecular Evolution 19, 255-271.

6. Casanova, M., Leroy, P., Boucekkine, C., Weissenbach, J., Bishop, C., Fellous, M., Purrello, M., Fiori, G., and Siniscalco, M. (1985). A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science 230, 1403-1406.

7. Wei, W., Ayub, Q., Chen, Y., McCarthy, S., Hou, Y., Carbone, I., Xue, Y., and Tyler-Smith, C. (2012). A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Research 23, 388-395.

8. Walker, A., and Smith, S. (1987). Mitochondrial DNA and human evolution. Nature 325, 1-5.

9. Forster, P. (2004). Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 359, 255-264.

10. Underhill, P.A., and Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet 41, 539-564.

11. Hammer, M.F., and Zegura, S.L. (2002). The human Y chromosome haplogroup tree: nomenclature and phylogeography of its major divisions. Annual Review of Anthropology, 303-321.

12. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., and Fritz, M.H.-Y. (2010). A draft sequence of the Neandertal genome. Science 328, 710-722.

13. Ghirotto, S., Tassi, F., Benazzo, A., and Barbujani, G. (2011). No evidence of Neandertal admixture in the mitochondrial genomes of early European modern humans and contemporary Europeans. American Journal of Physical Anthropology 146, 242-252.

14. Biswas, S., Scheinfeldt, L.B., and Akey, J.M. (2009). Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. The American Journal of Human Genetics 84, 641-650.

15. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2011). Inference of population structure using dense haplotype data. PLoS Genetics 8, e1002453.

16. Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., and Reich, D. (2012). A direct characterization of human mutation based on microsatellites. Nature genetics.

17. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., and Jonasdottir, A. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature 488, 471-475.

18. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., and Masson, G. (2002). A high-resolution recombination map of the human genome. Nature genetics 31, 241-247.

19. Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., and Akylbekova, E.L. (2011). The landscape of recombination in African Americans. Nature 476, 170-175.

20. Schaffner, S., and Sabeti, P. (2008). Evolutionary adaptation in the human lineage. Nature Education 1.

21. Manolio, T.A., and Collins, F.S. (2009). The HapMap and genome-wide association studies in diagnosis and therapy. Annual review of medicine 60, 443.

22. Novembre, J., and Ramachandran, S. (2011). Perspectives on Human Population Structure at the Cusp of the Sequencing Era. Annual review of genomics and human genetics 12, 245-274.

23. Tishkoff, S.A., Reed, F.A., Friedlaender, F.o.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., and Doumbo, O. (2009). The genetic structure and history of Africans and African Americans. Science 324, 1035-1044.

24. Johnson, N.A., Coram, M.A., Shriver, M.D., Romieu, I., Barsh, G.S., London, S.J., and Tang, H. (2011). Ancestral Components of Admixed Genomes in a Mexican Cohort. PLoS Genetics 7, e1002410.

25. Winkler, C.A., Nelson, G.W., and Smith, M.W. (2010). Admixture Mapping Comes of Age*. Annual review of genomics and human genetics 11, 65-89.

26. Chakraborty, R. (1986). Gene admixture in human populations: models and predictions. Yearbook Phys Anthropol 29.

27. Chakraborty, R., and Weiss, K. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proceedings of the National Academy of Sciences 85, 9119.

28. Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. Science 201, 786-792.

29. Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2, e190.

30. Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945.

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

32. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science 298, 2381-2385.

33. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A., Auton, A., Indap, A., King, K., Bergmann, S., and Nelson, M. (2008). Genes mirror geography within Europe. Nature 456, 98-101.

34. Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. Nature genetics 40, 646-649.

35. Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. The American Journal of Human Genetics 82, 290-303.

36. Price, A., Tandon, A., Patterson, N., Barnes, K., Rafaels, N., Ruczinski, I., Beaty, T., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genetics 5.

37. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genetics 8, e1002453.

38. 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. Human Biology 84, 343-364.

39. Pool, J., and Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181, 711.

40. Pugach, I., Matveyev, R., Wollstein, A., Kayser, M., and Stoneking, M. (2011). Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biol 12, R19.

41. Gravel, S. (2012). Population genetics models of local ancestry. Genetics 191, 607-619.

42. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Research 19, 318.

43. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., and Daly, M. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559-575.

44. Browning, B.L., and Browning, S.R. (2011). A fast, powerful method for detecting identity by descent. The American Journal of Human Genetics 88, 173-182.

45. Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P., Morrow, B., Friedman, E., Oddoux, C., and Burns, E. (2010). Abraham's Children in the Genome Era: Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry. The American Journal of Human Genetics.

46. Browning, S.R., and Browning, B.L. (2012). Identity by descent between distant relatives: detection and applications. Annual Review of Genetics 46, 617-633.

47. Cornuet, J.-M., Santos, F., Beaumont, M.A., Robert, C.P., Marin, J.-M., Balding, D.J., Guillemaud, T., and Estoup, A. (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics 24, 2713-2719.

48. Beaumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. Genetics 162, 2025-2035.

49. Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182, 1207-1218.

50. Nielsen, R., and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158, 885-896.

51. Becquet, C., and Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. Genome Research 17, 1505-1519.

52. Hey, J. (2005). On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS biology 3, e193.

53. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2010). Diffusion Approximations for Demographic Inference: DaDi.

54. Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. Nature genetics 43, 1031-1034.

55. Cavalli-Sforza, L.L., and Edwards, A.W. (1967). Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics 19, 233.

56. Thompson, E.A. (1975). Human evolutionary trees.(CUP Archive).

57. Lathrop, G. (1982). Evolutionary trees and admixture: phylogenetic inference when some populations are hybridized. Annals of human genetics 46, 245-255.

58. Waddell, P.J., and Penny, D. (1996). Evolutionary trees of apes and humans from DNA sequences. Handbook of symbolic evolution, 53-73.

59. Pickrell, J.K., and Pritchard, J.K. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8, e1002967.

60. Beerli, P., and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Sciences 98, 4563-4568.

61. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2013). Ancient admixture in human history. Genetics 192, 1065-1093.

62. Chakraborty, R., and Weiss, K.M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proceedings of the National Academy of Sciences 85, 9119-9123.

63. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. Genetics 192, 1065-1093.

64. Loh, P., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. (2013). Infering Admixture Histories of Human Populations using Weighted Linkage Disequilibrium. Genetics 193, 1233-1254.

65. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., and Mesa, N. (2012). Reconstructing Native American population history. Nature 488, 370-374.

66. Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., and Jombart, T. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334, 94-98.

67. Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. PLoS Genetics 8, e1002947.

68. Rife, D.C. (1954). Populations of hybrid origin as source material for the detection of linkage. American Journal of Human Genetics 6, 26.

69. Stephens, J.C., Briscoe, D., and O'Brien, S.J. (1994). Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. American Journal of Human Genetics 55, 809.

70. McKeigue, P.M. (1998). Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. The American Journal of Human Genetics 63, 241-251.

71. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O,ÄôBrien, S.J., and Altshuler, D. (2004). Methods for high-density admixture mapping of disease genes. American Journal of Human Genetics 74, 979.

72. Montana, G., and Pritchard, J.K. (2004). Statistical tests for admixture mapping with case-control and cases-only data. American Journal of Human Genetics 75, 771.

73. Haiman, C., Patterson, N., Freedman, M., Myers, S., Pike, M., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., and McDonald, G. (2007). Multiple regions within 8q24 independently affect risk for prostate cancer. Nature genetics 39, 638.

74. Choudhry, S., Taub, M., Mei, R., Rodriguez-Santana, J., Rodriguez-Cintron, W., Shriver, M.D., Ziv, E., Risch, N.J., and Burchard, E.G. (2008). Genome-wide screen for asthma in Puerto Ricans: evidence for association with 5q23 region. Human genetics 123, 455-468.

75. Consortium, T.S.T.D. (2013). An ancient haplotype carrying four missense SNPs in SLC16A11 is a common risk factor for type 2 diabetes in Mexico. Under review.

76. Dhandapany, P.S., Sadayappan, S., Xue, Y., Powell, G.T., Rani, D.S., Nallari, P., Rai, T.S., Khullar, M., Soares, P., and Bahl, A. (2009). A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. Nature genetics 41, 187-191.

Chapter 2

The History of African Gene Flow into Southern Europeans, Levantines and Jews*

Originally published as: Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.* 2011 Apr;7(4):e1001373. doi: 10.1371/journal.pgen.1001373. Epub 2011 Apr 21.

Author Contributions:

Conceived and designed the experiments: PM, NP, JNH, DR

Analyzed the data: PM, NP, DR

Contributed data: LH, GA, EB, HO, AK

Contributed analysis tools: PM, NP, ALP

Wrote the paper: PM, DR

2.1 Abstract

Previous genetic studies have suggested a history of sub-Saharan African gene flow into some West Eurasian populations after the initial dispersal out of Africa that occurred at least 45,000 years ago. However, there has been no accurate characterization of the proportion of mixture, or of its date. We analyze genome-wide polymorphism data from about 40 West Eurasian groups to show that almost all Southern Europeans have inherited 1-3% African ancestry with an average mixture date of around 55 generations ago, consistent with North African gene flow at the end of the Roman Empire and subsequent Arab migrations. Levantine groups harbor 4-15% African ancestry with an average mixture date of about 32 generations ago, consistent with close political, economic and cultural links with Egypt in the late middle ages. We also detect 3-5% sub-Saharan African ancestry in all eight of the diverse Jewish populations that we analyzed. For the Jewish admixture, we obtain an average estimated date of about 72 generations. We hypothesize that this may reflect descent of these groups from a common ancestral population that already had some African ancestry prior to the Jewish Diasporas.

2.2 Author Summary

Southern Europeans and Middle Eastern populations are known to have inherited a small percentage of their genetic material from recent sub-Saharan African migrations, but there has been no estimate of the exact proportion of this gene flow, or of its date. Here, we apply genomic methods to show that the proportion of African ancestry in many Southern European groups is 1-3%, in Middle Eastern groups is 4-15% and in Jewish groups is 3-5%. To estimate the dates when the mixture occurred, we develop a novel method that estimates the size of chromosomal

segments of distinct ancestry in individuals of mixed ancestry. We verify using computer simulations that the method produces useful estimates of population mixture dates up to 300 generations in the past. By applying the method to West Eurasians, we show that the dates in Southern Europeans are consistent with events during the Roman Empire and subsequent Arab migrations. The dates in the Jewish groups are older, consistent with events in classical or biblical times that may have occurred in the shared history of Jewish populations.

2.3 Introduction

The history of human migrations from Africa into West Eurasia is only partially understood. Archaeological and genetic evidence indicate that anatomically modern humans arrived in Europe from an African source at least 45,000 years ago, following the initial dispersal out of Africa[1; 2]. However, it is known that Southern Europeans and Levantines (people from modern day Palestine, Israel, Syria and Jordan) have also inherited genetic material of African origin due to subsequent migrations. One line of evidence comes from Y-chromosome[3] and mitochondrial DNA analyses[4-6]. These have identified haplogroups that are characteristic of sub-Saharan Africans in Southern Europeans and Levantines but not in Northern Europeans[7]. Auton et al.[8] presented nuclear genome-based evidence for sharing of sub-Saharan African ancestry in some West Eurasians, by identifying a North-South gradient of haplotype sharing between Europeans and sub-Saharan Africans, with the highest proportion of haplotype sharing observed in south/southwestern Europe. However, none of these studies used genome-wide data to estimate the proportion of African ancestry in West Eurasians, or the date(s) of mixture. Throughout this report, we use "African mixture" to refer to gene flow into West Eurasians since the divergence of the latter from East Asians; thus, we are not referring to the much older

dispersal out of Africa ~45,000 years ago but instead to migrations that have occurred since that time.

2.4 Results

We assembled data on 6,529 individuals drawn from 107 populations genotyped at hundreds of thousands of single nucleotide polymorphisms (SNPs). This included 3,845 individuals from 37 European populations in the Population Reference Sample (POPRES)[9; 10], 940 individuals from 51 populations in the Human Genome Diversity Cell Line Panel (HGDP-CEPH)[11; 12], 1,115 individuals from 11 populations in the third phase of the International Haplotype Map Project (HapMap3)[13], 392 individuals who self reported as having Ashkenazi Jewish ancestry from InTraGen Population Genetics Database (IBD)[14] and 237 individuals from 7 populations in the Jewish HapMap Project[15]. For most analyses, we used HapMap3 Utah European Americans (CEU) to represent Northern Europeans and HapMap3 Yoruba Nigerians (YRI) to represent sub-Saharan Africans, although we also verified the robustness of our inferences using alternative populations.

We curated these data using Principal Components Analysis (PCA)[16], with the most important steps being: (i) Removal of 140 individuals as outliers who did not cluster with the bulk of samples of the same group, (ii) Removal of all 8 Greek samples who separated into sub-clusters in PCA so that it was not clear which of these clusters was most representative, (iii) Splitting the Bedouins into two genetically discontinuous groups, and (iv) Reclassifying the 5 Italian groups into three ancestry clusters (Sardinian, Northern-Italy, and Southern-Italy) (see details in Text A.1, Figure A.1). A comparison of results before and after this curation is

presented in Table A.3, where we show that this data curation does not affect our qualitative inferences.

To study the signal of African gene flow into West Eurasian populations, we began by computing principal components (PCs) using South Africans (HGDP-CEPH- San) and East Eurasians (HapMap3 Han Chinese- CHB), and plotted the mean values of the samples from each West Eurasian population onto the first PC, a procedure that has been called "PCA projection"[17; 18]. The choice of San and CHB, which are both diverged from the West Eurasian ancestral populations[19; 20], ensures that the patterns in PCA are not affected by genetic drift in West Eurasians that has occurred since their common divergence from East Eurasians and South Africans. We observe that many Levantine, Southern European and Jewish populations are shifted towards San compared to Northern Europeans, consistent with African mixture, and motivating formal testing for the presence of African ancestry (Figure 2.1, Figure A.2).

To formally test for the presence of African mixture, we first performed the *4 Population Test*. This test is based on the insight that if populations *A* and *B* form sister groups relative to *C* and *D*, the allele frequency differences (p_A-p_B) and (p_C-p_D) should be uncorrelated as they represent independent periods of random genetic drift[21]. Applying the *4 Population Test* to the proposed relationship (YRI,(Papuan,(CEU,*X*))) where *X* is a range of West Eurasian populations, we find significant violations for all Southern European, Jewish and Levantine populations but not for Northern Europeans (Table 2.1). The results remain unchanged even when we use alternate topologies replacing YRI with other African populations (Text A.2, Table A.4). We further verified these inferences with the *3 Population Test*[21], which capitalizes on the insight that for any 3 populations (*X; A, B*), the product of the allele frequency differences (p_X-p_A) and (p_X-p_B) is expected to be negative only if population *X* descends from a mixture of populations

Table 2.1. Formal Tests for population mixture

Population (*X*)	Sam-ples	Region	Dataset	Z-score for *4 Pop. Test* $((P_x-P_{CEU}),(P_{Papuan}-P_{YRI}))$	Z-score for *3 Pop. Test* $((P_x-P_{CEU}),(P_x-P_{YRI}))$
African Americans	49	n/a	HapMap3	**-85.1**	**-108.9**
Palestine	43	L	HGDP-CEPH	**-27.9**	**-24.7**
Turkey	6	L	POPRES	-1	**-3.4**
Bedouin-g1	15	L	HGDP-CEPH	**-36**	**-40.7**
Bedouin-g2	30	L	HGDP-CEPH	**-25.8**	>0
Druze	41	L	HGDP-CEPH	**-14.6**	>0
Spain	137	SE	POPRES	**-12.3**	**-21.1**
Portugal	134	SE	POPRES	**-14.9**	**-29**
Romania	14	SE	POPRES	-0.5	**-5.1**
Croatia	6	SE	POPRES	0.7	>0
Bosnia-Herzegovina	9	SE	POPRES	-0.6	-1.5
Sardinia	27	SE	HGDP-CEPH	**-9.3**	>0
Southern-Italy	121	SE	POPRES	**-10.7**	**-14.2**
Northern-Italy	90	SE	POPRES	**-5.7**	**-5.7**
Austria	14	ECE	POPRES	-0.2	-2.4
Poland	22	ECE	POPRES	1.3	>0
Hungary	19	ECE	POPRES	0.4	**-5.6**
Czech Republic	11	ECE	POPRES	0.5	>0
Adygei	17	ECE	HGDP-CEPH	2.9	>0
Russia	6	ECE	POPRES	0.6	-0.2
Russia	25	ECE	HGDP-CEPH	11.4	>0
Swiss-French	759	I	POPRES	**-3.2**	**-6.1**
France	92	I	POPRES	-1.9	**-3.7**
France	28	I	HGDP-CEPH	-1.9	-2.9
Basque	24	I	HGDP-CEPH	-1.2	>0
Belgium	43	I	POPRES	-0.9	-2.2
Orkney	15	I	POPRES	3.2	>0
United Kingdom	388	I	POPRES	1.5	>0
Ireland	62	I	POPRES	1.7	>0
Scotland	5	I	POPRES	3.3	>0
Netherlands	17	I	POPRES	1.0	>0
Swiss-German	84	I	POPRES	-1	-2.6
Germany	74	I	POPRES	-0.9	-2.8
Sweden	11	I	POPRES	1.6	0
Ashkenazi Jews	323	n/a	IBD	**-11.6**	>0
Ashkenazi Jews	34	n/a	Jewish HapMap	**-9.5**	-2.2
Syrian Jews	25	n/a	Jewish HapMap	**-10.1**	-2.3
Iranian Jews	24	n/a	Jewish HapMap	**-5.9**	>0
Iraqi Jews	36	n/a	Jewish HapMap	**-8.5**	>0
Sephardic Greek Jews	39	n/a	Jewish HapMap	**-13.7**	**-15.2**
Sephardic Turkey Jews	27	n/a	Jewish HapMap	**-13.6**	**-17.1**
Italian Jews	27	n/a	Jewish HapMap	**-11.4**	>0

Notes: We analyzed data from all West Eurasian populations with ≥5 samples. Regions are abbreviated: I – Northwest Europe, ECE – East-Central Europe, SE – Southern Europe and L – Levant. We used a *Block Jackknife* (block size of 5cM) to correct for LD among SNPs and to estimate a *Z*-score that reports the number of approximately normally distributed standard deviations that the correlation coefficient differs from 0. For the *4 Population Test*, we interpret $|Z|>3$ as significant evidence for mixture (we test the tree $((P_x-P_{CEU})(P_{Papuan}-P_{YRI})$, and do not show the tests of the two alternative trees, although all $|Z|$-scores are >16). For the *3 Population Test,* we interpret $Z<-3$ as significant evidence for mixture; a positive score for the *3 Population Test* is possible even in the presence of population mixture, since genetic drift after mixture can mask the signal (for example, Bedouin-g2). Scores that are significant are highlighted in bold. For further study of sub-Saharan African mixture, we chose populations with a significantly negative score by the *4 Population Test* (bold).

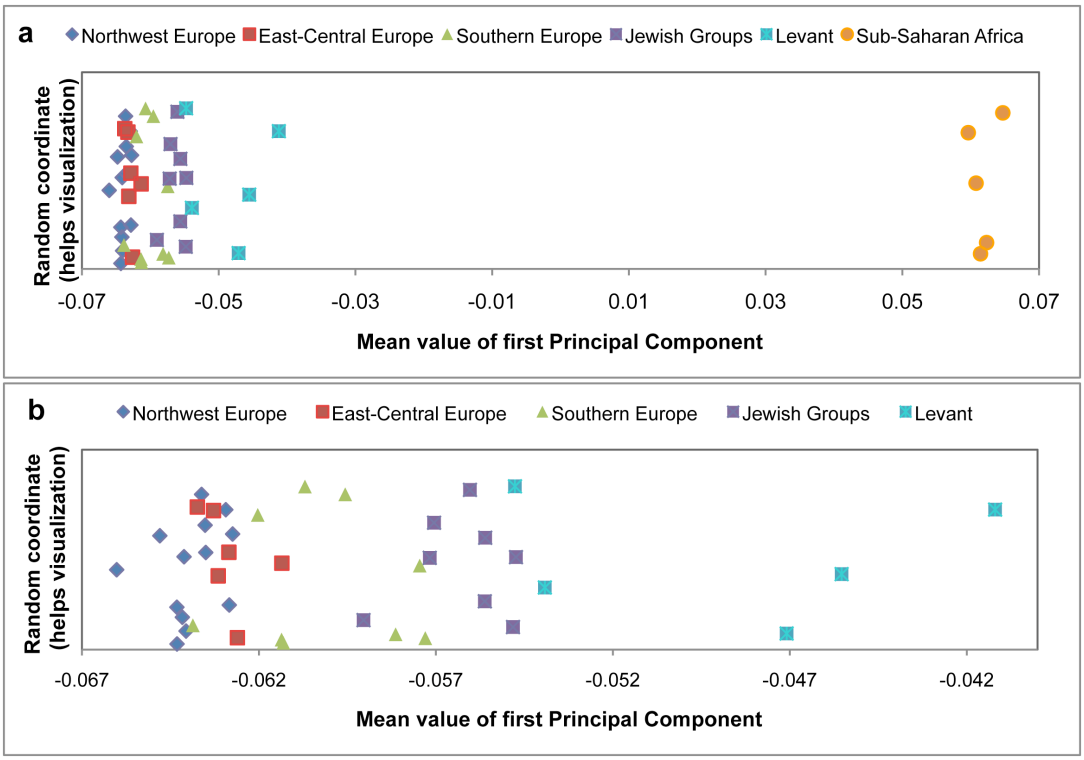


Figure 2.1 PCA Projection. PCA was performed using genome-wide SNP data from East Asians (HapMap3- CHB) and South Africans (HGDP-CEPH- San). All West Eurasians populations with samples sizes of $n \geq 5$ were then projected onto these PCs. (a) The first panel presents data for all populations and (b) the second panel provides a higher resolution view of West Eurasians after removing sub-Saharan Africans. Each point on this graph indicates the mean value of the first PC for a projected population. West Eurasians populations are colored by 5 regional groupings—"Northwest Europe", "East-Central Europe", "Southern Europe", "Levant", "Jewish Groups" (the assignments of populations to groups is shown in Table 2.1). The grouping "Sub-Saharan Africa" refers to six populations from the HGDP-CEPH panel: Kenyan Bantu, South African Bantu, Mandenka, Mbuti Pygmy, Biaka Pygmy and Yoruba.

related to populations *A* and *B*[21]. We verified that this method is robust to SNP ascertainment bias by carrying out simulations showing that the *3 Population Test* detects real admixture even if all SNPs used in the analysis are discovered in population *A*, population *B*, or in both

populations *A* and *B* (Text A.3). Application of the test to each West Eurasian population (using *A*=YRI and *B*=CEU) finds little or no evidence of mixture in North Europeans but highly significant evidence in many Southern European, Levantine and Jewish groups (Table 2.1).

To estimate the proportion of sub-Saharan African ancestry in the various West Eurasian populations that showed significant evidence of mixture, we used *f_4 Ancestry Estimation*[21], a method which produces accurate estimates of ancestry proportions, even in the absence of data from the true ancestral populations. This method estimates mixture proportions by fitting a model of mixture between two ancestral populations, followed by (possibly large) population-specific genetic drift. Briefly, we calculate a statistic that is proportional to the correlation in the allele frequency difference between West Eurasians and sub-Saharan Africans, and divide it by the same statistic for a population of sub-Saharan African ancestry, like YRI (Figure 2.2). This method has been shown through simulation to be robust to ascertainment bias on the SNP arrays and deviations from the assumed model of mixture (e.g. date and number of mixture events)[21].

Application of *f_4 Ancestry Estimation* suggests that the highest proportion of African ancestry in Europe is in Iberia (Portugal 3.2 ± 0.3% and Spain 2.4 ± 0.3%), consistent with inferences based on mitochondrial DNA[6] and Y chromosomes[7] and the observation by Auton et al.[8] that within Europe, the Southwestern Europeans have the highest haplotype-sharing with Africans. The proportion decreases to the north and we find no evidence for mixture in Russia, Sweden and Scotland (Table 2.2, Figure A.6). We also detect about 3-5% sub-African ancestry in all the Jewish populations, a finding that is novel as far as we are aware, and certainly has not been unambiguously demonstrated or quantified. For Levantines, the proportions are often higher: 9.3% ± 0.4% in Palestinians and >10% in the Bedouins (Standard errors were calculated using a Block Jackknife as described in Materials and Methods). Table 2.2 presents the ancestry

estimates that we obtain for all West Eurasian populations with significant evidence of mixture by the *4 Population Test* (*Z*-score < -3).

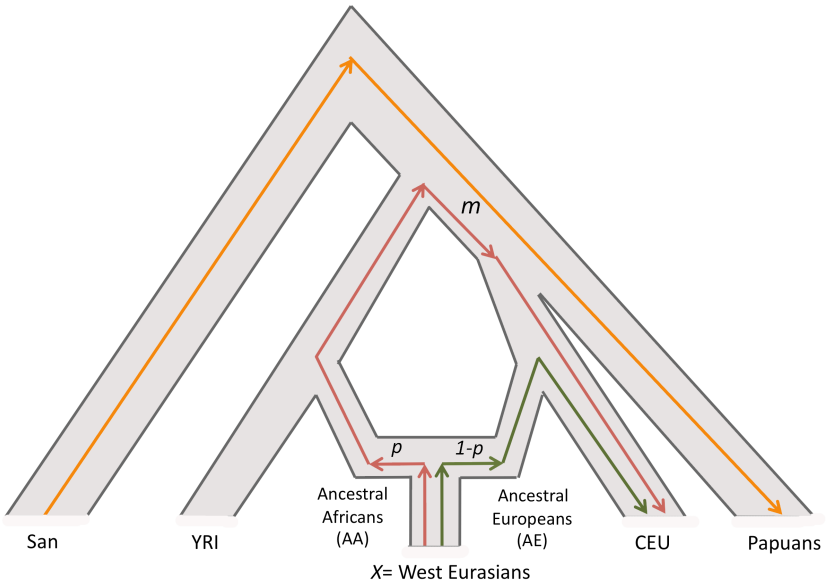


Figure 2.2 Estimation of African ancestry using *f_4 Ancestry Estimation*. *f_4 Ancestry Estimation* computes the quantity [(San-Papuan).(*X*–CEU)/ [(San-Papuan).(YRI-CEU)]; where *X* = any West Eurasian population. The denominator is proportional to the genetic drift *m* that occurred in the ancestors of West or East Africans since their divergence from San but prior to their divergence from West Eurasians (intersection of red and orange lines). The numerator is proportion to *p**(Ancestral Africans-YRI) + (1-*p*)*(Ancestral Europeans-CEU). Since the branches connecting (San, Papuan) and (CEU, *X*) do not overlap each other, the quantity (1-*p*)*(*X*-CEU) = 0 and hence the numerator is expected to equal *pm*. Thus, the ratio of the numerator and denominator is expected to equal *p* (Ancestral African mixture proportion). This figure is adapted from reference[21], where we first developed *f_4 Ancestry Estimation*, and where we reported computer simulations demonstrating its robustness.

Table 2.2. Estimates of mixture proportions and date of mixture

Population (X)	Dataset	Region	Sam -ples	West African ancestry proportion ± standard error	West African ancestry proportion using STRUCTURE	Estimated date of admixture (generations ± standard error)	Bias from simulations (generations)*	Estimated date of admixture after bias correction
African Americans	HapMap3	n/a	49	79.4% ± 0.3%	77.2%	6 ± 1	0	6 ± 1
Palestinian	HGDP-CEPH	L	43	9.3% ± 0.4%	11.0%	34 ± 2	1	33 ± 2
Bedouin-g1	HGDP-CEPH	L	15	14.5% ± 0.4%	15.6%	34 ± 3	2	32 ± 3
Bedouin-g2	HGDP-CEPH	L	30	10.1% ± 0.4%	11.6%	33 ± 2	2	31 ± 2
Druze	HGDP-CEPH	L	41	4.4% ± 0.4%	5.6%	54 ± 7	10	44 ± 7
Spain	POPRES	SE	137	2.4% ± 0.3%	1.1%	55 ± 3	0	55 ± 3
Portugal	POPRES	SE	134	3.2% ± 0.3%	2.1%	45 ± 5	0	45 ± 5
Sardinian	HGDP-CEPH	SE	27	2.9% ± 0.5%	0.2%	96 ± 28	25	71 ± 28
Southern-Italy	POPRES	SE	121	2.7% ± 0.3%	1.7%	62 ± 6	0	62 ± 6
Northern-Italy	POPRES	SE	90	1.1% ± 0.3%	0.2%	154 ± 27	-26	180 ± 27
Swiss-French	POPRES	I	759	0.5% ± 0.2%	0.1%	71 ± 6	n/a	n/a
Ashkenazi Jews	IBD	n/a	323	2.8% ± 0.3%	2.6%	91 ± 11	n/a	n/a
Ashkenazi Jews	Jewish HapMap	n/a	34	3.2% ± 0.4%	2.6%	76 ± 13	23	53 ± 13
Syrian Jews	Jewish HapMap	n/a	25	3.9% ± 0.5%	4.1%	99 ± 23	27	72 ± 23
Iranian Jews	Jewish HapMap	n/a	24	2.6% ± 0.6%	4.6%	129 ± 34	59	70 ± 34
Iraqi Jews	Jewish HapMap	n/a	36	3.8% ± 0.5%	4.5%	153 ± 22	38	115 ± 22
Sephardic Greek Jews	Jewish HapMap	n/a	39	4.8% ± 0.4%	3.7%	82 ± 8	20	62 ± 8
Sephardic Turkey Jews	Jewish HapMap	n/a	27	4.5% ± 0.4%	4.3%	89 ± 11	16	73 ± 11
Italian Jews	Jewish HapMap	n/a	27	4.9% ± 0.5%	4.0%	88 ± 19	15	73 ± 19

Note: Estimates of the proportions and dates of mixture for all populations that give statistically significant evidence of mixture in Table 2.1 (*4 Population Test* Z<-3). Regions are abbreviated as: I – Northwest Europe, SE – Southern Europe and L – Levant. Mixture proportion estimates are based on f_4 *Ancestry Estimation* using San, Yoruba, CEU and Papuan as the reference populations. The *ROLLOFF* estimated date of mixture uses CEU and YRI as the proposed ancestral populations (in the supplementary materials, we show that very similar inferences are obtained when the analysis is repeated with other ancestral populations, such as East Africans Luhya instead of Yoruba). Standard errors are computed using a Block Jackknife.

* Our simulations show that *ROLLOFF* produces a bias in the date estimates for small sample sizes, small mixture proportions, and old mixture dates. For each row of this table, we carried out a simulation to assess the expected bias for the inferred parameters (Table A.12) and we computed the bias as (average - true date) in generations. Based on the simulation results, we have corrected the estimate in the last column as (estimated date - bias). We do not report a correction for the two rows marked "n/a" because our simulator cannot accommodate this large sample size.

To test if our inferences are dependent on the sub-Saharan African population that was used as the reference group, we also repeated analyses with other sub-Saharan African populations replacing YRI. This analysis shows that our estimates of mixture proportions do not change significantly based on the ancestral population used (Text A.2c, Table A.5). We obtained similar estimates when we applied STRUCTURE 2.2[22] to estimate the mixture proportions using ~13,900 independent markers (that were not in linkage disequilibrium (LD) with each other) (Table 2.2, Figure A.5).

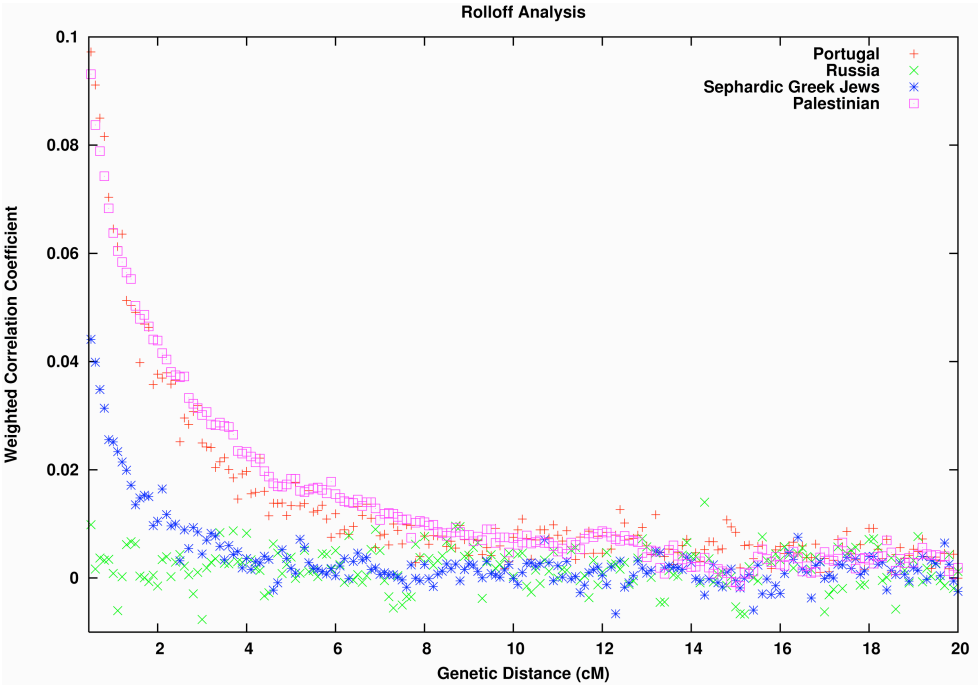


Figure 2.3 Testing for LD due to African admixture in West Eurasians. To generate these plots, we used the *ROLLOFF* software to calculate the LD between all pairs of markers in each population, weighted by their frequency difference between YRI and CEU to make the statistic sensitive to admixture LD. We plot the correlation as a function of genetic distance for Portuguese, Russians, Sephardic Greek Jews and Palestinians. We do not show inter-SNP intervals of <0.5cM since we have found that at this distance admixture LD begins to be confounded by background LD, and so inferences are not reliable (exponential curve fitting does not include inter-SNP intervals at this scale).

The finding of sub-Saharan African ancestry in West Eurasians predicts that there will be a signature of admixture LD in the populations that experienced this mixture. That is, there will be LD between all markers that are highly differentiated between the two ancestral populations and the allele will be strongly correlated to the local ancestry[23]. Hence, there will be chromosomal segments of African ancestry with lengths that reflect the number of recombination events that have occurred since mixture, and thus can be used to estimate an admixture date. Figure 2.3 shows that this expected pattern is observed empirically in the decay of LD in four example West Eurasian populations, where we enhance the effects of admixture LD by weighting the SNP comparisons by frequency difference between the ancestral Africans (YRI) and ancestral West Eurasians (CEU). In the Southern European, Jewish and Levantine populations, this procedure produces clear evidence of admixture LD (Figure 2.3). However, Northern Europeans (Russians in Figure 2.3) do not show any evidence of African gene flow, consistent with the *4 Population* and *3 Population Test* results and Figure 2.1. Similar results are seen for other West Eurasian and Jewish populations that show evidence of mixture in the *4 Population Test* (Figure A.10).

To estimate a date for the mixture event, we developed a novel method *ROLLOFF* that computes the time since mixture using the rate of exponential decline of admixture LD in plots such as Figure 2.3. *ROLLOFF* computes the correlation between a (signed) statistic for LD between a pair of markers and a weight that reflects their allele frequency differentiation in the ancestral populations. By examining the correlation between pairs of markers as they become separated by increasing genetic distance and fitting an exponential distribution to this rolloff by least squares, we obtain an estimate of

the date (see Materials and Methods and Text A.4). *ROLLOFF* also computes an approximately normally distributed standard error by carrying out Weighted Jackknife analysis [24], where we drop one chromosome in each run and study the fluctuation of the statistic in order to assess the stability of the estimate.

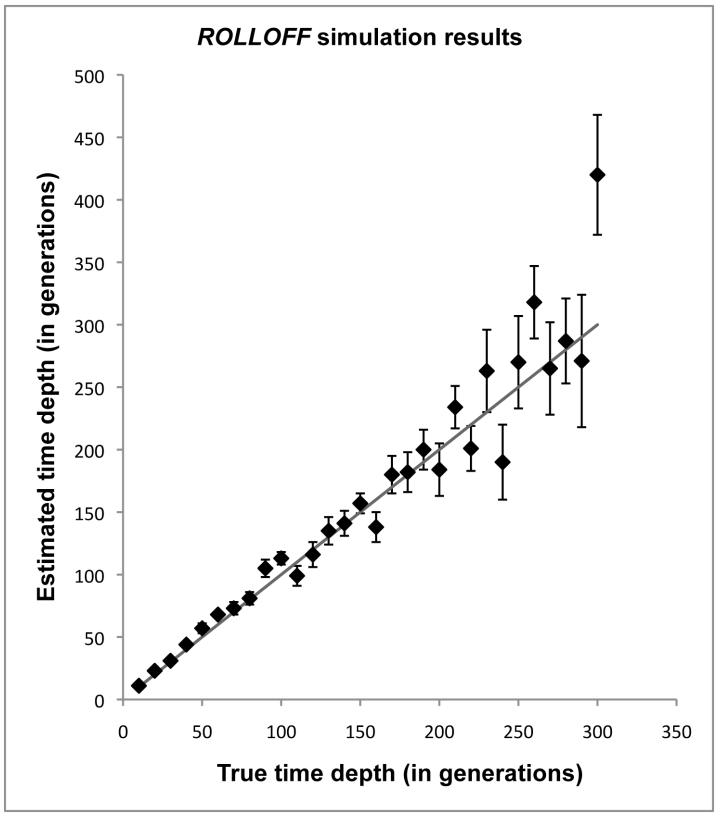


Figure 2.4 *ROLLOFF* simulation results. We constructed 10 individuals of mixed African and European ancestry (where individuals had 20% European ancestry) for various time depths ranging from 10-300 generations (with intervals of 10 generations). We performed *ROLLOFF* analysis using another independent dataset of European Americans and Nigerian Yoruba individuals as reference populations. We plot the true time depth (that was used for the simulations) against the estimated time depth computed by *ROLLOFF*. The expected time depth is shown as a dotted grey line. Standard errors were calculated using the *Weighted Block Jackknife* described in the Materials and Methods.

To verify the accuracy and sensitivity of *ROLLOFF*, we carried out extensive simulations by constructing the genomes of individuals of mixed ancestry by sampling haplotypes from North Europeans (CEU) and West Africans (YRI) (see Materials and Methods). We verified that *ROLLOFF* produces accurate estimates of the date of mixture, even in the case of old admixture (up to 300 generations – Figure 2.4) and is robust to substantially inaccurate ancestral populations as well as fine scale errors in the genetic map (Text A.4). In addition, to test the robustness of our inferences, we applied all the methods to African Americans and obtained consistent results for the proportion of mixture (79.4 ± 0.3%) and date of mixture (6 ± 1), which is in agreement with previous reports[25; 26]. However, in the case of low mixture proportion and old admixture dates, we observed that there is a slight bias in the estimated date (Text A.4d, Table A.9). This effect is related to the weakness of the signal: it attenuates as the sample size or admixture proportion becomes larger (Text A.4d, Table A.10, Table A.11).

An important concern was how *ROLLOFF* would perform when the true history of admixture involved multiple pulses of gene exchange, rather than the single pulse of gene exchange that we modeled. To explore this, we first simulated two distinct gene flow events, and then estimated the date using a single exponential distribution. The simulations show that *ROLLOFF*'s estimate of the date tends to correspond reasonably well to the more recent admixture event, with a slight upward bias towards the older date. Second, we performed simulations under a continuous gene flow model and found that the estimated dates are intermediate between the start and end of the gene flow, as expected. To see if we could obtain a better inference of the range of dates, we tried fitting sum of multiple exponential distributions, but this did not work reliably, which

may be related to the well-known difficulty of fitting a sum of exponentials to data with even a small amount of noise[27] (Text A.4). Pool and Nielsen recently showed that multi-marker haplotype data could be useful for distinguishing a single pulse of gene exchange from changing migration rates over time[28]. However, a complication with applying this approach to relatively old dates is that haplotype-based methods need to model background LD. In the case of old mixture events (dozens or hundreds of generations), inaccurate modeling of background LD can bias estimates[26; 29]. We are not aware of any published method that can produce accurate date estimates while modeling background LD correctly for mixture dates as old as those that have been explored by *ROLLOFF* in Figure 2.4.

We applied *ROLLOFF* to all the West Eurasian populations that gave significant signals of mixture by the *4 Population Test*, fitting a single exponential decay in each case. We estimate that the date of sub-Saharan African mixture in Portugal is 45 ± 5 generations and in Spain is 55 ± 3 generations. We estimate a more recent date of 34 ± 3 for Bedouin-g1, 33 ± 2 for Bedouin-g2, and 34 ± 2 generations for Palestinians. We estimate older dates of ~70-150 generations in the various Jewish populations, with wide and in most cases overlapping confidence intervals (Table 2.2, Figure A.10). Averaging the mixture dates over all populations from each region (weighted by the inverse of the squared standard error), we obtain an average of 55 generations for Southern Europeans, 34 for Levantines and 89 for Jews.

As described above, in our simulations to explore the behavior of *ROLLOFF* we detected an upward bias in the date estimates that grew worse with older mixture dates, small mixture proportions, and small sample sizes (but does not appear to be affected by

use of inaccurate ancestral populations). To assess the degree to which this bias might be affecting our date estimates, we performed simulations for each population in Table 2.2 separately, in which we set the number of samples, mixture proportion and time since mixture to match the parameters estimated from the real data. We repeated our simulations 100 times for each parameter setting and estimated the bias of our estimated date from the true (simulated) date. The bias is very small for the most of the Southern European and Levantine samples, which generally had large sample sizes, recent dates, and high mixture proportions. However, the bias is larger for the Jewish groups (Table 2.2, Table A.12). Correcting for the bias inferred in our simulation of Table A.12, we obtain corrected estimates of the average date of 55 generations for Southern Europeans, 32 for Levantines, and 72 for Jews. A caveat about these regional date estimates is that they reflect weighted averages across the populations in each region. However, it is important to recognize that the admixture events detected within each region may not reflect the same historical events; for example, it is plausible that the sub-Saharan African admixture in Spain and Italy have different historical origins.

2.5 Discussion

The finding of African ancestry in Southern Europe dating to ~55 generations ago, or ~1,600 years ago assuming 29 years per generation[30], needs to be placed in historical context. The historical record documents multiple interactions of African and European populations over this period. One potential opportunity was during the period of Roman occupation of North Africa that lasted until the early 5th century AD, and

indeed tomb inscriptions and literary references suggest that trade relations continued even after that time[31; 32]. North Africa was also a supplier of goods and products such as wine and olive oil to Italy, Spain and Gaul from 200-600 AD, and Morocco was a major manufacturer of the processed fish sauce condiment, garum, which was imported by Romans[33]. In addition, there was slave trading across the western Sahara during Roman times[7; 34]. Another potential source of some of the African ancestry, especially in Spain and Portugal, is the invasion of Iberia by Moorish armies after 711 AD[35; 36]. If the Moors already had some African ancestry when they arrived in Southern Europe, and then admixed with Iberians, we would expect the admixture date to be older than the date of the invasion, as we in fact observe.

The signal of African mixture that we detect in Levantines (Bedouins, Palestinians and Druze) – an average of 32 generations or ~1000 years ago – is more recent than the signal in Europeans, which might be related to the migrations between North Africa and Middle East that have occurred over the last thousand years, and the proximity of Levantine groups geographically to Africa. Syria and Palestine were under Egyptian political control until the 16th century AD when they were conquered by the Ottoman Empire. This is in concordance with our proposed dates. In addition, the Arab slave trade is responsible for the movement of large numbers of people from Africa across the Red Sea to Arabia from 650 to 1900 AD and probably even prior to the Islamic times[7; 37]. We caution that our sampling of the Middle East is sparse, and it will be of interest to study African ancestry in additional groups from this region.

A striking finding from our study is the consistent detection of 3-5% sub-Saharan African ancestry in the 8 diverse Jewish groups we studied, Ashkenazis (from northern

Europe), Sephardis (from Italy, Turkey and Greece), and Mizrahis (from Syria, Iran and Iraq). This pattern has not been detected in previous analyses of mitochondrial DNA and Y chromosome data[7], and although it can be seen when re-examining published results of STRUCTURE-like analyses of autosomal data, it was not highlighted in those studies, or shown to unambiguously reflect sub-Saharan African admixture[15; 38]. We estimate that the average date of the mixture of 72 generations (~2,000 years assuming 29 years per generation[30]) is older than that in Southern Europeans or other Levantines. The point estimates over all 8 populations are between 1,600-3,400 years ago, but with largely overlapping confidence intervals. It is intriguing that the Mizrahi Irani and Iraqi Jews—who are thought to descend at least in part from Jews who were exiled to Babylon about 2,600 years ago[39; 40] —share the signal of African admixture (An important caveat is that there is significant heterogeneity in the dates of African mixture in various Jewish populations). A parsimonious hypothesis for these observations is that they reflect a history in which many of the Jewish groups descend from a common ancestral population which was itself admixed with Africans, prior to the beginning of the Jewish diaspora that occurred in 8th to 6th century BC[41]. The dates that emerge from our *ROLLOFF* analysis in the non-Mizrahi Jews could also reflect events in the Greek and Roman periods, when there were large communities of Jews in North Africa, particularly Alexandria[34; 42]. We note that we detect a similar African mixture proportion in the non-Jewish Druze (4.4 ± 0.4%) although the date is more recent (54 ± 7 generations; 44 ± 7 after the bias correction). Algorithms such as PCA and STRUCTURE show that various Jewish populations cluster with Druze[15], which coupled with the similarity in mixture proportions, is consistent with descent from a common ancestral population. Importantly,

the other Levantine populations (Bedouins and Palestinians) do not share this similarity in the African mixture pattern with Jews and Druze, making them distinct in their admixture history.

A caveat to these results is that we estimated dates assuming instantaneous mixture, but in fact we have not distinguished between the patterns expected for instantaneous admixture and continuous gene flow over a long period. In Text A.4f, we report simulations showing that for continuous gene flow, the dates from *ROLLOFF* reflect the average of mixture dates over a range of times, and so the date should be interpreted only as an average number.

A potential issue that could in theory influence our findings is that the exact population contributing to African ancestry in West Eurasians is unknown. To gain insight into the African source populations, we carried out PCA analyses, which suggested that the African ancestry in West Eurasians is at least as closely related to East Africans (e.g. Hapmap3 Luhya (LWK)) as to West Africans (e.g. Nigerian Yoruba (YRI)) (the same analyses show that there is no evidence of relatedness to Chadic populations like Bulala) (Text A.5). We also used the *4 Population Test* to assess whether the tree ((LWK, YRI),(West Eurasian, CEU)) is consistent with the data, and found no evidence for a violation, which is consistent with a mixture of either West African or East African ancestors or both contributing to the African ancestry in West Eurasians (Table A.14). Historically, a mixture of West and East African ancestry is plausible, since African gene flow into West Eurasia is documented from both West Africa during Roman times[34] and from East Africa during migrations from Egypt[7]. It is important to point out, however, that the difficulty of pinpointing the exact African source population

is not expected to bias our inferences about the total proportion and date of mixture. The *f_4 Ancestry Estimation* method is unbiased even when we use a poor surrogates for the true ancestral African population (as long as the phylogeny is correct), as we confirmed by repeating analyses replacing YRI with LWK, and obtaining similar results (Table A.15). Our *ROLLOFF* admixture date estimates are also similar whether we use LWK or YRI to represent ancestral African population (Table A.15), as predicted by the theory.

In summary, we have documented a contribution of sub-Saharan African genetic material to many West Eurasian populations in the last few thousand years. A key area for future work should be to identify the source populations for this admixture, which should provide new insights about West Eurasian history.

2.6 Materials and Methods

Datasets

We analyzed individuals of West Eurasian ancestry from several sources: The Population Reference Sample (POPRES)[9; 10] (n=3,845 samples from 37 populations genotyped on an Affymetrix 500K array), the Human Genome Diversity Cell Line Panel (HGDP-CEPH)[12] (n=940 samples from 51 populations genotyped on an Illumina 650K array), The International Haplotype Map (HapMap) Phase 3[13] (n=1,115 samples from 11 populations genotyped on an Illumina 1M array), the InTraGen Population Genetics Database (IBD)[14] (n=392 Ashkenazi Jews genotyped on an Illumina 300K array) and the Jewish HapMap Project[15] (n=232 from 7 Jewish populations genotyped on an Affymetrix 6.0 array). We created a merged dataset containing 6,529 individuals -out of which 3,614 individuals of West Eurasian, African and Eastern Eurasian ancestry were used for the final analysis.

Detailed information about the number of individuals and markers included in each analysis is provided in Table A.1. We used NCBI Build 35 genetic map to determine physical position and the Oxford LD-based map to determine genetic positions of all SNPs[43].

Methods for characterizing mixture

Principal Component analysis (PCA): PCA was performed using *smartpca,* part of the EIGENSOFT 3.0 package[16]. For the PCA Projection analysis, the *poplistname* flag was used to compute Principal Components (PCs) on only a subset of populations from the dataset[17; 18]. The merged dataset M with 36,175 SNPs was used for this analysis (Table A.1).

4 Population Test: For any 4 populations (*A*, *B*, *C*, *D*), there are three possible unrooted phylogenetic trees. If the tree ((*A*, *B*), (*C*, *D*)) is correct, then the genetic drift separating *A* and *B* should not be correlated to the drift separating *C* and *D*. However, if mixture occurred, then the correlation might be non-zero (Figure A.3). We compute the correlation as in reference[21], and use a *Block Jackknife* [24; 44] that drops 5 centimorgan (cM) blocks of the genome in each run, to compute a standard error of the statistic. We convert the correlation into a *Z*-score and test for mixture by assessing whether the *Z*-score is more than 3 standard deviations different from 0. To test for sub-Saharan African mixture in West Eurasians, we tested the unrooted phylogenetic tree ((YRI,Papuan),(CEU,*X*)) where *X* is a range of West Eurasian populations. For this analysis, we intersected the HGDP-CEPH and HapMap3 data with all other datasets (POPRES, IBD, Jewish HapMap) to preserve the maximum number of SNPs. The

merged datasets G, J, K and L with ~606K, ~85K, ~284K and ~118K SNPs respectively were used for these analyses (Table A.1).

3 Population Test: The *3 Population Test* can verify if population *X* is related to populations *A* and *B* through a simple tree or has arisen due to mixture. For a simple tree, the product of the frequencies differences between *A* and *X*, and *B* and *X*, is expected to be positive[21]. We compute a *Z*-score reporting the number of standard deviations that the statistic differs from 0, using the same Block Jackknife procedure as described above. A significantly negative value provides an unambiguous signal for mixture in *X* related to populations *A* and *B*[21] (also see Figure A.3). For this analysis, we intersected HapMap3 dataset individually with all other datasets (HGDP-CEPH, POPRES, IBD, Jewish HapMap). The merged datasets F, G, H, I containing ~347K, ~606K, ~284K and ~466K SNPs respectively were used for the analysis (Table A.1).

f_4 Ancestry Estimation: We assume the population relationships shown in Figure 2.2 and denote the allele frequency of SNP *i* in each population as $p_{San}{}^{i}$, $p_{Papuan}{}^{i}$ $p_{YRI}{}^{i}$ $p_{CEU}{}^{i}$ and $p_{X}{}^{i}$ (*X* = any West Eurasian population). To estimate the proportion of sub-Saharan African ancestry in population *X*, we compute the ratio of two *4 Population Test* statistics:

$$f_4(San,YRI;CEU,Papuan) = \frac{\sum_{i=1}^{n}(p_{San}i - p_{Papuan}i)(p_{X}i - p_{CEU}i)}{\sum_{i=1}^{n}(p_{San}i - p_{Papuan}i)(p_{YRI}i - p_{CEU}i)}$$

This quantity is summed over all markers and the standard errors are computed using the Block Jackknife[24; 44] (block size of 5cM). The numerator is proportional to the amount of

sub-Saharan African-related ancestry in population *X*, while the denominator is the same quantity for a population of entirely sub-Saharan African ancestry (YRI). Thus, the ratio estimates the mixture proportion[21] (Figure 2.2). The merged datasets G, J, K and L with ~606K, ~85K, ~284K and ~118K SNPs respectively were used for this analysis (Table A.1).

STRUCTURE 2.2: To obtain an independent estimate of mixture proportions, we applied the model based clustering algorithm implemented in STRUCTURE 2.2 [22] to all populations that showed evidence of admixture using *4 Population Test* (Table 2.1). As a control, we also added HapMap3 African Americans (ASW) and two Northern European populations, Russia and Sweden. To make the run tractable, we reduced the number of markers to 13,877 SNPs. We excluded all the SNPs that were in LD with other SNPs in a window of 0.1cM. We ran STRUCTURE without any prior population assignment (unsupervised mode), with K = 2 and with 10,000 iterations for burn-in and 10,000 follow-on iterations. We used the INFERALPHA option under the admixture model.

Estimating the date of admixture

Overview of ROLLOFF: To estimate dates of ancient admixture, we have developed a method, *ROLLOFF*, which examines pairs of SNPs and assesses how admixture related LD decreases with genetic distance. The method is based on a novel LD statistic that weights SNPs according to their allele frequency differentiation between two populations that are genetically 'close' to the ancestral mixing populations.

Suppose that we have an admixed population and for simplicity assume that the population is homogeneous and that the mixture occurred over a short time span, ideally only a few generations. Call the two admixing populations A, B, and suppose that the admixture event occurred n generations before the present. If we consider two SNPs that are at a distance d Morgans apart on a chromosome in an admixed individual, then with probability e^{-nd} the alleles at these SNPs derived from a single admixing individual. If the mixing proportions are p_A and p_B respectively ($p_A + p_B = 1$), then we see that:

1. With probability $e^{-nd}p_A$, both alleles belong to population A.
2. With probability $e^{-nd}p_B$ both alleles belong to population B.
3. With probability $(1-e^{-nd})$ the alleles belong to populations A or B independently.

We next suppose that we have a weight function at each SNP that is positive when the variant allele is more likely to be in population A than B and negative in the reverse situation. If $w(s)$ is the weight of SNP s, then for any pair of SNPs s_1, s_2, we aim to compute an LD-based score $z(s_1,s_2)$ that is asymptotically standard normal and positive if the two variant alleles are in admixture LD. As we explain below, the score $z(s_1,s_2)$ and the product of the weight functions $w(s_1)\cdot w(s_2)$ are expected to be correlated, and to have a correlation coefficient exactly proportional to e^{-nd}.

To convert the z-scores between all possible pairs of SNPs into an estimate of mixture age, we bin the z-scores based on the distance separations d, and compute the correlation coefficient between $z(s_1,s_2)$ and $w(s_1)\cdot w(s_2)$ in each bin. Fitting an exponential distribution to the fall-off of the correlation coefficient with distance, we compute the admixture date from the fitted exponent. Our simulations show that the optimal bin size is

at least 0.05cM; smaller bins result in very short inter-SNP intervals so that analysis becomes confounded by background LD. In practice, we use a bin size of 0.1cM.

Mathematical details of the ROLLOFF weight function: If we have data from two populations *A'* and *B'* that are genetically close to the admixing populations, then if *a*, *b* are the empirical allele frequencies at an allele for a SNP *s* in the two populations, we propose the weight function $w(s) = (a-b)/\sqrt{p(1-p)}$ where $p = (a+b)/2$. A valuable feature of our *ROLLOFF* method is that we can also calculate useful weights even when no suiTable A.urrogate parental populations are available (making it impossible to obtain direct estimates of the ancestral allele frequencies), by simply choosing a weight function that is proportional to the allele frequency difference, even if the absolute values cannot be computed directly.

Mathematical details of the ROLLOFF LD score $z(s_1,s_2)$: To compute an LD score $z(s_1,s_2)$ for two SNPs s_1 and s_2 we use the following procedure:

1. We compute the Pearson correlation coefficient ρ for the diploid genotypes at s_1 and s_2. Samples with missing data at either marker are ignored. Let N be the number of samples with non-missing data. Setting $z = \sqrt{N}\rho$ would probably be satisfactory but we slightly refine this. We insist that $N \geq 4$.

2. We 'clip' ρ to fall within the interval [-0.9, 0.9].

3. We set $x = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$, which is Fisher's z-transformation.

4. We finally set $z(s1,s2) = \sqrt{N-3}x$

If the 2 markers (s_1,s_2) are unlinked, then z is roughly standard normal because of Fisher's *z*-transformation. Note that if the markers are unlinked, no matter how z is

defined, our weight function will be uncorrelated. This suggests that our method is robust to any reasonable definition of z.

Estimation of Standard Errors: We implemented a *Weighted Block Jackknife Test* [24,44] where we drop one chromosome in each run and study the fluctuation of the statistic in the 22 runs. The statistic estimated in each run is weighted by the number of SNPs excluded in that run. By studying the variability of the estimated date, we compute the uncertainty in the inferred quantity via the theory of the jackknife [24]. These standard errors should be viewed with some caution as they reflect only 22 independent outcomes.

The reason we have chosen to carry out the jackknife on the scale of an entire chromosome is that we are concerned that LD due to admixture may extend sufficiently far for some populations that jackknifing by much smaller blocks (e.g. 10 Mb) may not completely remove the correlation among segments. We have therefore taken a conservative approach and set the block sizes to be equal to a chromosome. However, for a key West Eurasian population (Spain), we repeated the analysis with block sizes of 5cM, 10cM and 20cM, as well as whole chromosomes and observed that the standard errors are similar (Table A.16).

Simulation framework to test ROLLOFF: We simulated individuals of mixed European and African ancestry such that the genome of each individual is a mosaic of haplotypes from both the ancestral populations. The method we used is adapted from the simulation method that we previously described in reference[26]. Briefly, our simulations are based on two parameters: (a) the mixture proportion (θ) that gives the probability that a particular sampled haplotype comes from European or African gene pool, and (b) the time of mixture (λ) which can be viewed as the number of generations since mixture. We

jointly phased data for 113 CEU individuals and 107 YRI individuals using fastPHASE[45] to create an ancestral haplotype pool of 226 haploid CEU and 214 haploid YRI genomes, which served as the source data for our simulations.

To simulate the genome of an admixed individual, we start at the beginning of each chromosome and sample European haplotypes with probability (θ) and African haplotypes with probability ($1-\theta$). At each marker, we resample ancestry with probability of $1-e^{-\lambda g}$ where g is the genetic distance in Morgans to determine if an event has occurred and then resample ancestry based on θ. Once the ancestry is chosen, a chromosomal segment of a randomly picked individual of that ancestry is then copied to the genome of the admixed individual and the process is continued until the end of chromosome is reached. This procedure is repeated to create the genomes of 20 admixed individuals, taking care that no chromosomal segment is reused (sampling without replacement). We combined pairs of haploid individuals to construct 10 diploid admixed individuals. This algorithm has one limitation that it requires more than $2n$ ancestral haplotypes for generating data for n diploid admixed individuals. Hence, in cases when we needed to simulate data for $n \geq 50$, we made a slight modification to the algorithm such that each admixed haploid genome is constructed from one haploid CEU and one haploid YRI genome, without reusing any chromosomal segments.

In order to test the performances for *ROLLOFF* at varying time depths, we performed 30 simulations. In each simulation, we constructed 10 diploid genomes of individuals of mixed European and African ancestry where we set λ = 10, 20...300 (interval = 10 generations) and θ = 20%. We performed *ROLLOFF* analysis (for each of the simulations) using a non-overlapping dataset of 1,107 European American and 737

Nigerian Yoruba individuals as reference samples to compute the allele frequency in the ancestral populations. All analyses were restricted to 339,171 SNPs and the fine scale recombination map by Myers et al.[43] was used for mapping the genetic distance.

ROLLOFF analysis of West Eurasian populations: We ran *ROLLOFF* for various West Eurasian populations using the HapMap3 CEU and YRI as reference populations. The correlation between SNPs was plotted as a function of genetic distance. To estimate a date, we fitted an exponential distribution to the decay of the correlation coefficients. The merged datasets F, G, H, I with ~347K, ~606K, ~284K and ~466K SNPs respectively were used for this analysis (Table A.1).

Software: ROLLOFF software can be downloaded as part of ADMIXTOOLS package from http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html

Acknowledgements

We are grateful to Philip DeJager and Richard Cooper for sharing the data for European Americans and Yoruba individuals that were used for estimating the allele counts needed for the *ROLLOFF* simulations. We are also grateful to Amy Williams, Noah Zaitlen and Bogdan Pasaniuc for allowing us to use their *simulator* software that was used for generating data under the continuous admixture model. We thank Michael McCormick and Kyle Harper for discussions about the historical contexts for these findings.

2.7 References

1. Stringer, C., and Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. Science 239, 1263-1268.

2. Prugnolle, F., Manica, A., and Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. Current Biology 15, R159-R160.

3. Adams, S., Bosch, E., Balaresque, P., Ballereau, S., Lee, A., Arroyo, E., López-Parra, A., Aler, M., Grifo, M., and Brion, M. (2008). The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. The American Journal of Human Genetics 83, 725-736.

4. Côrte-Real, H., Macaulay, V., Richards, M., Hariti, G., Issad, M., Cambon-Thomsen, A., Papiha, S., Bertranpetit, J., and Sykes, B. (1996). Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Annals of human genetics 60, 331-350.

5. Dupanloup, I., Bertorelle, G., Chikhi, L., and Barbujani, G. (2004). Estimating the impact of prehistoric admixture on the genome of Europeans. Molecular biology and evolution 21, 1361.

6. Amorim, A., Alves, C., Cunha, C., and Pereira, L. (2005). African female heritage in Iberia: a reassessment of mtDNA lineage distribution in present times. Human Biology 77, 213-229.

7. Richards, M., Rengo, C., Cruciani, F., Gratrix, F., Wilson, J., Scozzari, R., Macaulay, V., and Torroni, A. (2003). Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. The American Journal of Human Genetics 72, 1058-1064.

8. Auton, A., Bryc, K., Boyko, A., Lohmueller, K., Novembre, J., Reynolds, A., Indap, A., Wright, M., Degenhardt, J., and Gutenkunst, R. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Research 19, 795.

9. Nelson, M., Bryc, K., King, K., Indap, A., Boyko, A., Novembre, J., Briley, L., Maruyama, Y., Waterworth, D., and Waeber, G. (2008). The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. The American Journal of Human Genetics 83, 347-358.

10. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A., Auton, A., Indap, A., King, K., Bergmann, S., and Nelson, M. (2008). Genes mirror geography within Europe. Nature 456, 98-101.

11. Rosenberg, N., Pritchard, J., Weber, J., Cann, H., Kidd, K., Zhivotovsky, L., an d Feldman, M. (2002). Genetic structure of human populations. In., pp 2381-2385.

12. Li, J., Absher, D., Tang, H., Southwick, A., Casto, A., Ramachandran, S., Cann, H., Barsh, G., Feldman, M., and Cavalli-Sforza, L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100.

13. Altshuler, D., Brooks, L., Chakravarti, A., Collins, F., Daly, M., and Donnelly, P. (2005). A haplotype map of the human genome. Nature 437, 1299-1320.

14. Mitchell, M., Gregersen, P., Johnson, S., and Parsons, R. (2004). The New York Cancer Project: rationale, organization, design, and baseline characteristics. Journal of Urban Health 81, 301-310.

15. Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P., Morrow, B., Friedman, E., Oddoux, C., and Burns, E. (2010). Abraham's Children in the Genome Era: Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry. The American Journal of Human Genetics.

16. Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2, e190.

17. McVean, G. (2009). A Genealogical Interpretation of Principal Components Analysis.

18. Patterson, N., Petersen, D., van der Ross, R., Sudoyo, H., Glashoff, R., Marzuki, S., Reich, D., and Hayes, V. (2009). Genetic structure of a unique admixed population: implications for medical research. Human molecular genetics.

19. Sun, J., Mullikin, J., Patterson, N., and Reich, D. (2009). Microsatellites are molecular clocks that support accurate inferences about history. Molecular biology and evolution 26, 1017.

20. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science 324, 1035-1044.

21. Reich, D., Thangaraj, K., Patterson, N., Price, A., and Singh, L. (2009). Reconstructing Indian population history. Nature 461, 489-494.

22. Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945.

23. Chakraborty, R. (1986). Gene admixture in human populations: models and predictions. Yearbook Phys Anthropol 29.

24. Busing, F., Meijer, E., and Leeden, R. (1999). Delete-m Jackknife for Unequal m. Statistics and Computing 9, 3-8.

25. Smith, M., Patterson, N., Lautenberger, J., Truelove, A., McDonald, G., Waliszewska, A., Kessing, B., Malasky, M., Scafe, C., and Le, E. (2004). A high-density admixture map for disease gene discovery in african americans. The American Journal of Human Genetics 74, 1001-1013.

26. Price, A., Tandon, A., Patterson, N., Barnes, K., Rafaels, N., Ruczinski, I., Beaty, T., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genetics 5.

27. Osborne, M., and Smyth, G. (1986). An algorithm for exponential fitting revisited. Journal of Applied Probability, 419-430.

28. Pool, J., and Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181, 711.

29. Falush, D., Stephens, M., and Pritchard, J. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567.

30. Fenner, J. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology 128, 415.

31. Boardman, J., Griffin, J., and Murray, O. (2001). The Oxford history of the Roman world.(Oxford University Press, USA).

32. Harris, W. (1980). Towards a study of the Roman slave trade. Memoirs of the American Academy in Rome 36, 117-140.

33. Curtis, R. (2005). Sources for Production and Trade of Greek and Roman Processed Fish. Ancient Fishing and Fish Processing in the Black Sea Region, 31-46.

34. Gibbon, E. (1890). The decline and fall of the Roman Empire.(WW Gibbings).

35. Kennedy, H. (1996). Muslim spain and portugal.(Longman).

36. O'Callaghan, J. (1983). A history of medieval Spain.(Cornell Univ Pr).

37. Segal, R. (2001). Islam's Black slaves.(Farrar, Straus and Giroux).

38. Behar, D., Metspalu, E., Kivisild, T., Achilli, A., Hadid, Y., Tzur, S., Pereira, L., Amorim, A., Quintana-Murci, L., and Majamaa, K. (2006). The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. The American Journal of Human Genetics 78, 487-497.

39. Levy, H., and Ebrami, H. (1999). Comprehensive history of the Jews of Iran.(Mazda Publ.).

40. Rejwan, N. (1985). The Jews of Iraq: 3000 years of history and culture.(Westview Press).

41. Stillman, N. (1979). The Jews of Arab lands: a history and source book.(Jewish Publication Society).

42. Ashtor, E. (1992). The Jews of Moslem Spain.(Jewish Publication Society of America).

43. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. In. (American Association for the Advancement of Science), pp 321-324.

44. Kunsch, H. (1989). The jackknife and the bootstrap for general stationary observations. The Annals of Statistics, 1217-1241.

45. Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics 78, 629-644.

Chapter 3

Indian Siddis: African descendants with Indian admixture*

*Originally published as: Shah AM[+], Tamang R[+], Moorjani P[+], Rani DS, Govindaraj P, Kulkarni G, Bhattacharya T, Mustak MS, Bhaskar LV, Reddy AG, Gadhvi D, Gai PB, Chaubey G, Patterson N, Reich D, Tyler-Smith C, Singh L, Thangaraj K. (2011) Indian Siddis: African descendants with Indian admixture. *Am J Hum Genet*. 2011 Jul 15;89(1):154-61. doi: 10.1016/j.ajhg.2011.05.030. Epub 2011 Jul 7.

[+]Authors contributed equally

Author Contributions:

Conceived and designed the experiments: KT, LS

Analyzed the data: *Mitochondrial and Y-Chromosome Analysis*- AMS, RT, GC, DSR, PG, KT, CTS; *Autosomal Analysis*- PM, NP, DR

Contributed data: AMS, RT, PG, DSR, TB, GK, MSM, LVSB, AGR, DG, PBG

Contributed analysis tools: PM, NP

Wrote the paper: PM, DR, AMS, RT, GC, KT

3.1 Abstract

The Siddis (Afro-Indians) are a tribal population who live in coastal Karnataka, Gujarat and in some parts of Andhra Pradesh. Historical records indicate that the Portuguese brought the Siddis to India from Africa about 300-500 years ago; however, there is little information about their more precise ancestral origins. Here, we perform a genome-wide survey to understand the population history of the Siddis. Using hundreds of thousands of autosomal markers, we show that they have inherited ancestry from Africans, Indians and possibly Europeans (Portuguese). Additionally, analyses of the uniparental (Y-chromosome and mitochondrial DNA) markers indicate that the Siddis trace their ancestry to Bantu speakers from sub-Saharan Africa. We estimate that the admixture between the Africans ancestors of the Siddis and neighboring South Asian groups likely occurred in the past 8 generations (~200 years ago), consistent with historical records.

3.2 Introduction

Siddis or Habshis are a unique tribe with African ancestry who live in South Asia. They are mainly found in three Indian states- Gujarat, Karnataka and Andhra Pradesh – and according to the latest census their total population size is about 0.25 million.[1] The first documented record of Siddis in India goes back to 1100 AD, when the Siddis settled in Western India.[2,3] By the thirteenth century, substantial numbers of Siddis were being imported by the Nawabs and the Sultans of India to serve as soldiers and slaves. The major influx of Siddis occurred during the 17th to 19th centuries when the Portuguese brought them as slaves to India.[2] Previous genetic studies have shown that the Siddis have ancestry from up to three continental groups: Africans, Europeans and South Asians.[2,4,5] Some genetic studies have suggested that they are most closely

related to Africans.[3,6] However, the specific African group that the Siddis trace their ancestry to remains unknown. To obtain a high-resolution genome-wide perspective of ancestry, we analyzed data from three Siddi groups (from Karnataka and Gujarat) by genotyping them with ~850,000 autosomal and sex-linked markers. Applying statistical methods, we have estimated the contributions of various continental ancestries to the Siddis genome, and investigated the likely source of the ancestral populations and the timing of the admixture events.

Blood samples (about 10ml from each individual) were collected from Gujarat and Karnataka states in India. Specifically, we collected samples from 59 Siddis (unrelated and healthy males) and 90 individuals belonging to the nearby tribal populations (Charan and Bharwad) of the Junagarh district of Gujarat state and from 94 Siddis (65 males and 29 females) and 178 individuals belonging to neighboring tribal populations (Medar, Gram Vokkal, Kare Vokkal and Korova) from the Uttara Kannad district of Karnataka state. Informed written consent was obtained from all the donors. This project was approved by the Institutional Ethical Committee (EIC) of the Centre for Cellular and Molecular Biology (CCMB), Hyderabad, India. We genotyped 16 Siddi samples on Affymetrix (SNP 6.0) arrays using standard protocols. We removed four duplicate samples and restricted the analysis to all the SNPs that have < 5% missing data (846,418 SNPs). We removed 4 duplicate samples and restricted the analysis to all the SNPs that have < 5% missing data (846,418 SNPs). There were two merged datasets created for further analysis: *Dataset-I*: Siddi data was merged with data from: The International Haplotype Map (HapMap) Phase 3[7] (n=1,115 samples from 11 populations genotyped on Illumina 1M array) and India Project (n = 132 individuals from 25 groups genotyped on an Affymetrix 6.0 array). The merged dataset contained 574,197 SNPs. Dataset-I was used for the Principal component analysis reported in Figure 3.1. *Dataset-II:* The Siddi data was merged with

three other datasets: The Population Reference Sample (POPRES)[8] (n=3,845 samples from 37 populations genotyped on an Affymetrix 500K array), The International Haplotype Map (HapMap) Phase 3[7] (n=1,115 samples from 11 populations genotyped on Illumina 1M array) and India Project (n = 132 individuals from 25 groups genotyped on an Affymetrix 6.0 array). The merged dataset contained 257,840 SNPs. Dataset-II was using for the estimating admixture proportions and dates of admixture.

3.3 Results

To explore patterns of population structure in the Siddis and to test their genetic affinity to other groups worldwide, we analyzed autosomal data from 12 individuals from three Siddi groups (6 individuals from Karnataka and 6 individuals from Gujarat), 128 individuals from 16 Indian groups (Mala, Madiga, Kurumba, Bhil, Kamsali, Satnami, Vysya, Naidu, Lodi, Tharu, Velama, Srivastava, Meghaval, Vaish, Kashmiri Pandit and Hallaki) and 300 individuals from three HapMap populations (Yoruba from Ibadan, Nigeria (YRI), Utah residents with Northern and Western European ancestry (CEU) and Han Chinese from Beijing, China (CHB)).[7,8] The 16 Indian groups were chosen to span a high degree of diversity within India. It had been previously shown that most Indian populations have ancestry from two highly divergent groups: an Ancestral North Indian (ANI) population which is closely related to West Eurasians and an Ancestral South Indian (ASI) population, which is not related to any population outside India. The ANI ancestry proportion lies within the range of 39-71% across the 16 groups chosen.[7] The ANI and ASI have been inferred to be highly differentiated at the time that they mixed and in Reich et al (2009)[7], it was estimated that the average allele frequency differentiation F_{ST}(ANI, ASI) is ~0.09.

We performed Principal Component Analysis (PCA) using the EIGENSOFT software[10] on the autosomal SNP data. A plot of the first and second Principal Components (PCs) suggests that the Siddis have ancestry from Africans as well as Eurasians (Figure 3.1a). Like other Indian populations, Siddis have both ANI and ASI ancestry but they lie off the main cline of ANI-ASI admixture and are closely related to African individuals (Figure 3.1a). The average allele frequency differentiation between the two Siddi groups (Karnataka and Gujarat) is relatively high: F_{ST} (Siddi_Karnataka-1, Siddi_Gujarat) = 0.02 (Table B.1), suggesting that the populations differ substantially, possibly due to endogamy or different ancestral origins or admixture with different local South Asian groups. However, the diversity in the Siddis is not correlated with geography in our small sample, as the individuals from the Karnataka2 group are genetically closer to the Gujarat Siddis (F_{ST} (Siddi_Karnataka-2, Siddi_Gujarat) = 0.002) than to the other group from Karnataka (F_{ST} (Siddi_Karnataka-1, Siddi_Karnataka2) = 0.026) (Table B.1). This suggests that the Karnataka2 samples may be recent migrants from Gujarat, or that the ancestors of one of the Karnataka samples may have experienced a very strong recent founder event.

Previous genetic studies using traditional biochemical and autosomal markers have suggested that the Siddis have ancestry from up to three distinct ancestral groups- Africans, Indians and European[3,6]. To formally test if the Siddis have ancestry from each of three ancestral populations (Africans, Indians and Europeans), we used a regression method by Patterson et al. (2010)[11] that models the allele frequency of the admixed Siddis as a linear combination of the allele frequencies in the ancestral populations. This method allows us to build optimal models with and without each ancestral population and then compute the error between our model and the data. For example, to test if the Siddis contain genetic admixture from Africans, we built two models using the data - one that included Africans as the ancestral population and another

excluded Africans from the model. Applying this method to the Siddi Gujarat samples, we observed that there is strong evidence that the Siddis have African ancestry (Z-score >> 27) but the genetic variation in Africans does not fully explain the underlying genetic data in the Siddis (Table B.2). Next, we compared if a two-way model or three-way mixture model provides a good fit to the data.

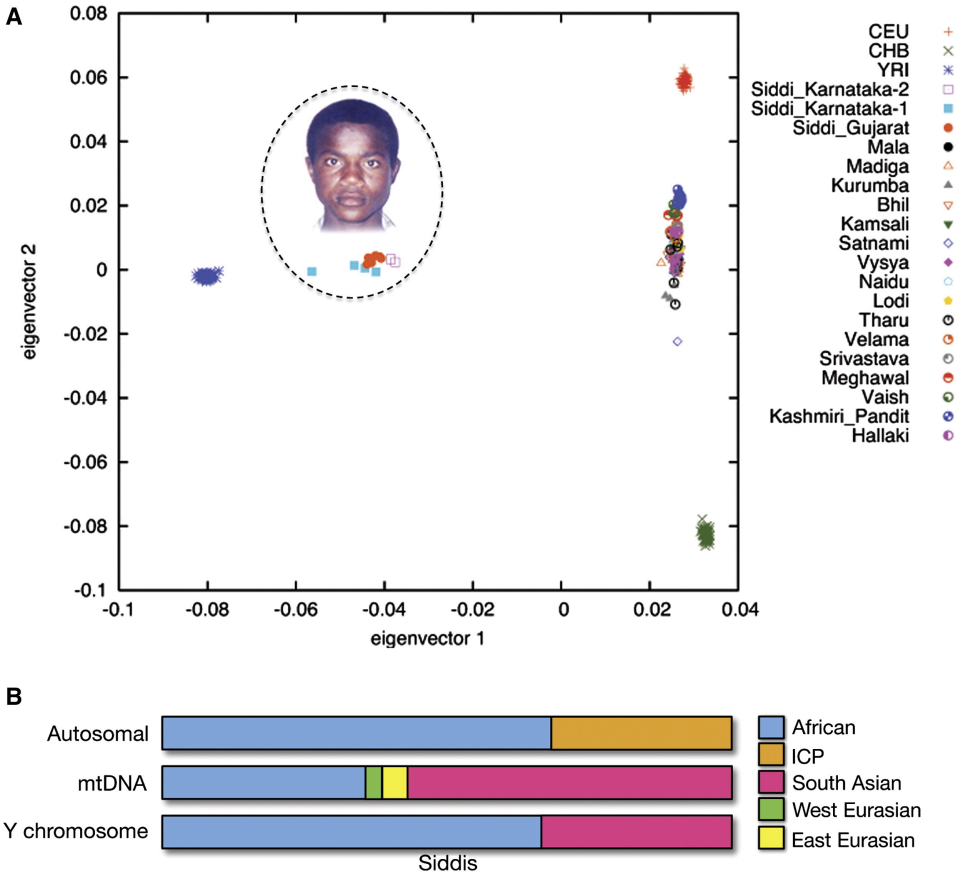


Figure 3.1 African ancestries in Siddis. a. Principal Component analysis of three Siddi groups, HapMap Phase 3 populations (CEU, YRI, CHB) and 16 Indian groups; **b**. Schematic representation of the proportions of African and ICP ancestry in Siddis. ICP (combined 16 Indian groups and Portuguese), which represents the ancestral non-African population.

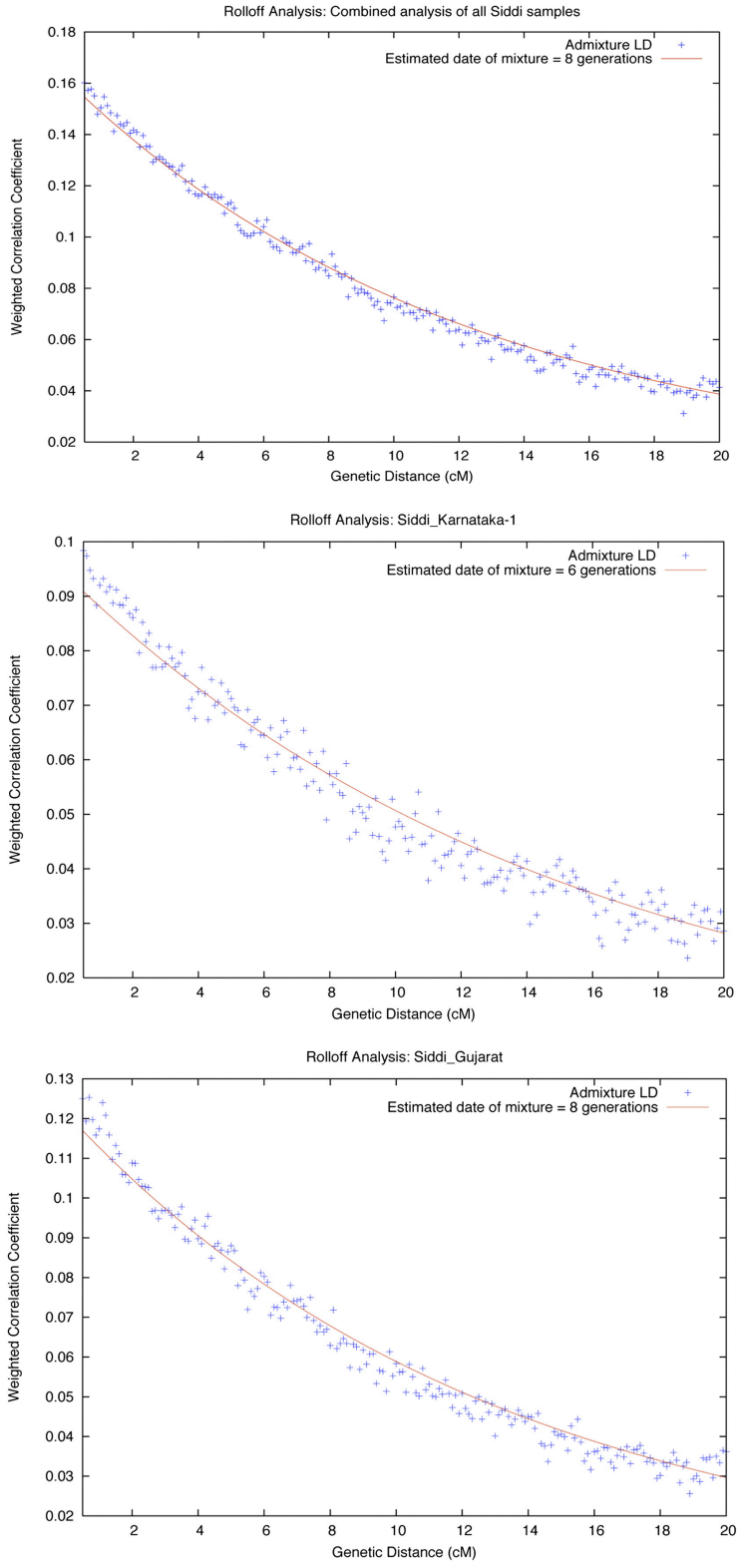


Figure 3.2 *ROLLOFF* analysis of Siddis. We analyzed 12 Siddi samples from Karnataka and Gujarat and estimated admixture linkage disequilibrium by computing the LD between all pairs

Figure 3.2 (Continued) of markers and weighting it by the frequency differentiation between the ancestral populations (YRI and ICP). We observed an approximately exponential decay of LD with distance with an average estimated date of admixture of 8 ± 1 generations. This corresponds to a time of around 200 years (assuming a generation interval of 25 years). The estimated dates of admixture for Siddi_Karnataka-1 and Siddi_Gujarat are 6 ± 1 and 8 ± 1 generations, respectively. Standard errors were computed using a Weighted Block Jackknife as described in reference [13].

Table B.3 shows that a two-way model of African and Portuguese or African + Mala (or any other group that has high ASI ancestry) provides a poor fit to the data. However, the model of African + Vaish (or any other group that has high ANI ancestry) provides just as good fit to the data as a three-way model of African + any Indian population + Portuguese. This suggests that the Siddis have some West Eurasian related (ANI or Portuguese) ancestry, in addition to their African and ASI ancestry. In addition, our methods are not sensitive enough to differentiate between ANI and Portuguese ancestry given our data set size. To represent the ancestral non-African population of the Siddis, we combined the data from 16 Indian groups and Portuguese ("ICP"). To test the robustness of our models, we analyzed Siddi Karnataka samples with the models built from the Siddi Gujarat samples and showed that the models provided a good fit to the data (Table B.4).

Applying the regression style method to all three Siddi groups with YRI and ICP as the ancestral populations, we estimated that the Siddis have on average ~67% African ancestry (Table 3.1). We obtained qualitatively similar results when we use East Africans (HapMap Luhya (LWK)) in place of YRI (Table 3.1 & Table B.4).

Table 3.1. **Estimation of ancestry proportions in the Siddis**

African ancestral population = West Africans (YRI)		
	African ancestry	Non-African ancestry (ICP)
Siddi_Gujarat	66.90% ± 0.59%	33.10% ± 0.59%
Siddi_Karnataka-1	70.90% ± 0.65%	29.10% ± 0.65%
Siddi_Karnataka-2	62.30% ± 0.99%	37.70% ± 0.99%
African ancestral population =East Africans (LWK)		
	African ancestry	Non-African ancestry (ICP)
Siddi_Gujarat	70.50% ± 0.66%	29.50% ± 0.66%
Siddi_Karnataka-1	74.40% ± 0.71%	25.60% ± 0.71%
Siddi_Karnataka-2	64.80% ± 1.11%	35.20% ± 1.11%

NOTE: Admixture proportion estimates are based on Regression style method[11] using the ancestral populations shown in the table. ICP - combined data from 16 Indian groups and Portuguese- represents the ancestral non-African population.

To characterize the temporal impact of admixture and to develop a historical interpretation of the results, it is important not only to qualitatively demonstrate a history of admixture, but also to quantitatively estimate a date for the admixture event. We applied the *ROLLOFF* method[12], which utilizes information related to admixture linkage disequilibrium (LD) to estimate the time since admixture. This method capitalizes on the insight that the genome of an admixed population contains chromosomal segments from ancestral populations, whose length is inversely proportional to the date of admixture. By modeling the decay of the LD in the admixed individuals and weighting it by the allele frequency differentiation in the ancestral populations (such that the statistic is only sensitive to admixture LD), we can precisely estimate the time since the admixture event. Simulations have suggested that this method is robust to data from poor surrogates of ancestral populations and can estimate the date of admixture up to 300 generations ago.[12]

Applying *ROLLOFF* to the Siddis (data from all three groups- Siddi_Karnataka-1, Siddi_Karnataka-2 and Siddi_Gujarat were combined to increase power), we observed an

approximately exponential decay of the weighted correlation with distance, which provides strong evidence of admixture (Figure 3.2). By fitting an exponential distribution to this pattern using least squares, we estimated an average date of ~8 generations or 200 years (assuming a generation size of 25 years[13]). This approximately coincides with the historical date of arrival of most African ancestors of the Siddis to India. To show that combining the data from the admixed group does not substantially change the results, we ran *ROLLOFF* separately for each admixed group and obtained qualitatively similar results (within two standard errors) for Siddi_Gujarat and Siddi_Karnataka-1. Due to limited number of samples, we were not able to perform analysis for the Siddi_Karnataka-2 group (*ROLLOFF* analysis requires at least 4 samples). In addition, changing the Africans ancestral group to East African Luhya did not change the estimated date of admixture (Figure B.1).

To gain insight into the most likely source of the African ancestry in Siddis, we examined paternally inherited Y-chromosomal biallelic markers as well as maternally inherited mtDNA markers. Analysis of data from uniparentally-inherited markers can provide information about population genetic relatedness, including probable ancestral source populations and information related to admixture events. We genotyped 32 Y-chromosomal biallelic markers (viz. M94, M60, M182, M168, M130, M145, M96, M75, M2, M89, M82, M304, M172, M9, M70, M11, M45, M207, M173, M17, M124, M201, M170, M70, M147, M189, M214, M52, M33, M356, P36 and P2) in 125 Siddis and 268 individuals (all males) from nearby Indian groups. We combined our data with published data from 2,301 individuals belonging to 56 different groups from the African subcontinent and 667 individuals from 16 populations from Gujarat, Karnataka, Maharashtra and Andhra-Pradesh states of India (Supporting Dataset S1).[14-26]

We observed that the Y-chromosomal haplogroups B2-M182 and E1b1a-M2 which are characteristic of African ancestry were present at high frequencies in the Siddis but not in other Indians. Moreover, about 70% of the Siddi male lineages fall into haplogroups generally characteristic of African populations (Figure 3.3a), thus confirming the results from the autosomal DNA markers (Figure 3.1e). The remaining 30% were C*-M130_ and M89-derived Indian or Near-Eastern lineages (H1a-M82, H2-Apt-H2, J2-M172, L-M11 and P*-M45). The populations neighboring the Siddis were found to harbor only these Asian-specific haplogroups. It is interesting to note that while none of the African paternal lineages were observed among the neighboring Indian groups, while Indian-specific lineages were detected in Siddi individuals. This suggests primarily unidirectional paternal gene flow from Indian populations to the Siddis (Figure 3.2b).

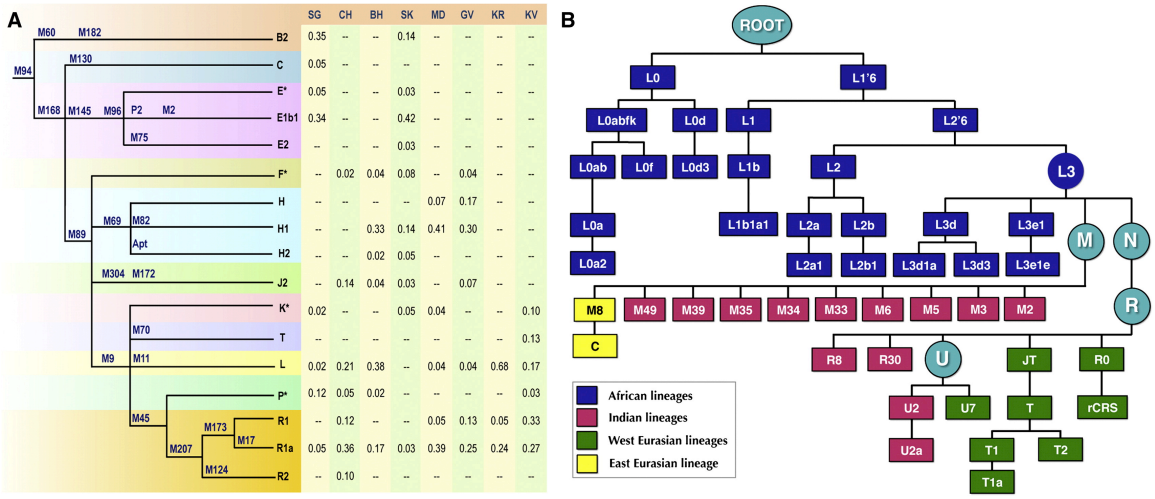


Figure 3.3 Y-chromosomal and mtDNA haplogroups in Siddis. a. Y-chromosomal haplogroup frequencies in the populations analyzed. SG; Siddis from Gujarat, CH; Charan, BH; Bharwad, SK; Siddis from Karnataka, MD; Medar, GV; Gram Vokkal, KR; Korova and KV; Kare Vokkal. The African-specific B; and E1b1a haplogroups were found in India exclusively among the Siddi population. **b**. Distribution of mtDNA haplogroups in Siddis. Details of diagnostic mutations which define haplogroups are shown in Supplementary Figure B.5, which define the haplogroups.

To learn more about the source of the African paternal lineages, we performed PCA with a merge of our Y-chromosomal dataset (Siddis and neighboring Indian groups) with data from 2,301individuals from 56 African populations (Supporting Dataset S1). A plot of the first and second PCs showed that the Siddis cluster with Bantu-speaking populations of sub-Saharan Africa (Figure B2a). Previous studies have proposed that the E3a (currently known as E1b1a), E2 and B2 haplogroups are associated with the Bantu expansions within Africa.[21,22,27] The presence of these haplogroups in the Siddis suggests that their ancestors may have been part of this expansion. To investigate this possibility, We typed 17 Y-STR using multiplex PCR using Y-filer® kit (Applied Biosystems, Foster City, USA) in reaction volumes of 10µl with 1U of AmpliTaq Gold® DNA polymerase (Applied Biosystems, Foster City, USA), 10mM Tris-HCl (pH 8.3), 50mM KCl, 1.5 mM $MgCl_2$, 250µM dNTPs, 3.0µM of each primer (forward primers are fluorescent labeled) and 1ng of DNA template. Thermal cycling conditions were as follows: (1) 95°C for 11 min, (2) 30 cycles: 94°C for 1 min, 61°C for 1 min, 72°C for 1 min, (3) 60°C for 80 min, and (4) 25°C hold. The PCR amplicons along with GS500 LIZ (as size standard) were run in the ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, USA). The raw data was analyzed using the GeneMapper v4.0 software program (Applied Biosystems, Foster City, US).

Two DYS385 loci were excluded from the current analyses because they could not be distinguished using the typing method employed and locus DYS389I was renamed as DYS389b, while DY389a was calculated by subtracting DYS389I from DYS389II.We constructed median-joining networks with 10 common loci (Figure B3) for the two major African haplogroups (E1b1a-M2 and B2-M182) that are present at high frequencies in Siddis. We supplemented our dataset with other published data that included African samples. [28,29] The TMRCA (time to most recent common ancestor) was estimated using the ρ statistic (the mean number of mutations from

the assumed root), using a 25-year generation time, and the TD statistic (both assuming a mutation rate of 6.9 x 10^{-4}per STR per generation),[30] The majority of the Siddis haplotypes were found shared on otherwise Bantu-specific branches and were present all over the tree (Figure B3). In addition, the Gujarat and Karnataka Siddis were highly diverged and did not share any haplotypes. These results support the autosomal observation of high *F*st differentiation among Siddis from Gujarat and Karnataka. Although the majority of the Siddis haplotypes were scattered in the network, we found that all haplogroup B2 Gujarat Siddis formed a cluster and coalesced to their most recent common ancestor 2.4±1 Kya (thousand years ago). The sharing of haplotypes suggests relatedness among the samples. This is similar to the results seen in the autosomal analyses of the Siddi_Gujarat and Siddi_Karnataka-2 samples. The male effective population size was estimated using BATWING[31,32,] with a demographic model that assumes a period of constant size followed by exponential growth (the prior probabilities for the other parameters used in the model were set as previously described).[26] A random subset of 40 samples was analyzed using 10^6 to 10^8 MCMC cycles and we obtained the same posterior probability for effective population size (N) as that obtained for 10^7 cycles. The effective population size of the African ancestors of Siddis brought to India during the slave trade was estimated as ~1,400 individuals (Table B.3 & Figure B4).

To gain insight into the maternal lineages and to test the directionality of maternal gene flow in the Siddis, we assayed the hypervariable region I (HVRI) of mtDNA in 153 Siddis and 269 individuals from the nearby Indian populations (accession numbers JN022021–JN022442). This data was compared with revised Cambridge Reference Sequence (rCRS)[33] for scoring the variations. Haplogroups were assigned based on HVR1 variations and they were further

confirmed by genotyping the coding regions mutations published till date (www.phylotree.org). The mtDNA haplogroup distributions in the Siddis are shown in Supporting Dataset 2.

Haplogroups were assigned using HVR-I (hypervariable region-I) variations, and were further confirmed with coding region variants (Figure 3.3b & Figure B5) (www.phylotree.org.).[34] PCA plots of the combined dataset (Supporting Dataset S2) showed that the African-specific mtDNA haplogroups were present at high frequency in the Siddis, similar to the observations from the autosomal and paternal lineages (Figure B2b). The African-specific haplogroup L was present at a frequency of 53% and 24% in Siddis from Gujarat and Karnataka, respectively. Previous studies have suggested that the L0a, L2a, L3b and L3e haplogroups are associated with the Bantu expansion.[35-39] Haplogroup L2a (including L2a1) was observed in the Siddis along with rare sublineages of L2, which further supports the conclusion that the ancestors of the Siddis were most likely African Bantus (Figure B5). The L0d lineage now largely confined to the Khoisan-speaking South Africans populations, but possibly more widespread in the past[40], was also observed in two Siddis individuals from Gujarat state. The presence of Indian-specific sub-lineages of M and N (R and U) (that include M2, M3, M5, M6, M33, M35, M39, M57, R8, R30 and U2 haplogroups) is indicative of recent admixture with indigenous Indian populations (Figure B5).[26] In addition, haplogroup T which is widespread in southern and Western Europe[41] and also present at low frequency in some South Asian groups[42] was present among four Siddi individuals (Figure B5). This may suggest maternal gene flow from a West Eurasian ancestral source - perhaps Portuguese or ANI. Consistent with the Y-chromosomal results, there is no evidence of African haplogroups in the neighboring Indian populations, thus, confirming the hypothesis of unidirectional gene flow to Siddi individuals from contemporary Indian populations (Figure 3.1b).

In order to further explore the evidence of sub-Saharan ancestry, we analyzed data for the *G6PD* (MIM 305900) variants in Siddis along with 26 ethnic populations from India. The A- variant, which provides protection against malarial infection and is estimated to have a sub-Saharan African origin between 3,840 to 11,760 YBP,[43] was observed only in Siddis (10%) and not in any other Indian populations (Table B.4). This further strengthens the evidence for the sub-Saharan ancestry of the Siddis.

3.4 Discussion

In conclusion, our combined analysis of genetic variation in the Siddis, using high-resolution sex-linked and autosomal markers, provides strong evidence of African ancestry together with unidirectional gene flow from local Indian groups to the Siddis. The directionality of gene flow supports the complex genetic structuring among Indian populations, which are highly influenced by social norms. We have traced the likely ancestry of Indian Siddis to sub-Saharan African Bantus. The ancestry proportions based on the analysis of autosomal and Y-chromosomal markers are similar while mtDNA markers reveal more South Asian lineages among Siddi individuals. The model that emerges from our results is as follows: During the course of the Bantu expansion, African farmers settled in East Africa. Later, during the 15th to 17th centuries, this region was predominantly ruled by the Portuguese. They brought some Africans to India as slaves and sold them to local Nawabs and Sultans, whose descendents, admixed with neighboring populations, comprise the present-day Siddi population of India (Figure 3.4).

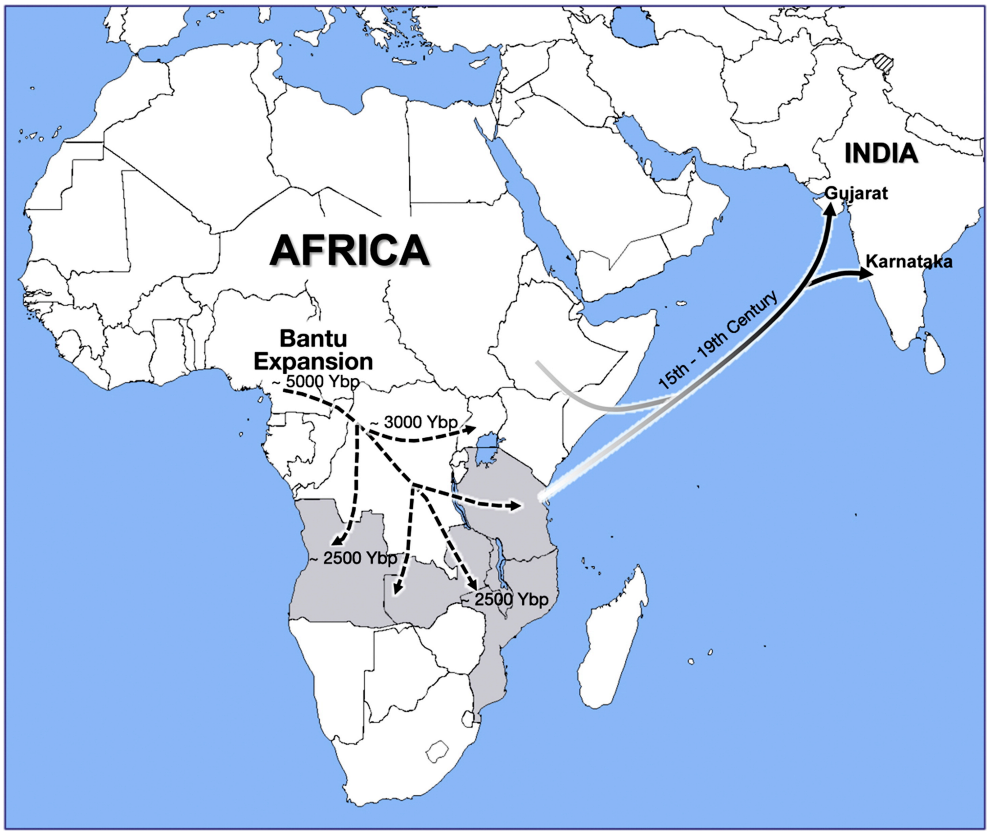


Figure 3.4 Migration history of the Siddis. Dotted arrows represent the expansion of Bantu-speakers with agriculture within Africa, which started from central western Africa and proceeded towards the east and south of the African subcontinent. The shaded grey area represents the Portuguese territory and the lines between Africa and India represent the possible path used by Portuguese during the 15th – 19th centuries to supply African slaves to Indian rulers on the western coast of India.

3.5 Web Resources

Programs:

The URLs for data presented herein are as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.omim.org

Phylotree, http://www.phylotree.org/

ROLLOFF software can be downloaded as part of ADMIXTOOLS package from:

http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html

Supporting Dataset S1 & S2: http://www.cell.com/AJHG/supplemental/S0002-9297(11)00223-0

Accession Numbers:

The sequences of the mtDNA hypervariable region I (HVRI) from 153 Siddis and 269 indivdiuals from the nearby Indian populations were deposited in the GenBank database under accession numbers JN022021–JN022442.

Acknowledgements

We thank all the donors, who have voluntarily donated their blood samples for this study. KT was supported by UKIERI (RG-4772), ICMR and CSIR, Govt. of India. CTS was supported by The Wellcome Trust and GC was supported by the Centre of Excellence of Estonian Biocentre. DR and PM were supported by the National Science Foundation. LS was supported by Bose Fellowship, DST, Govt. of India and Bhatnagar fellowship, CSIR, Government of India.

3.6 References

1. Lodhi, A., (1992). African settlements in India. Nordic journal of African studies. *1*, 83-86.

2. Bhattacharya, D., (1970). Indians of African origin. Cahiers d'Etudes Africaines *10*, 579-582.

3. Gauniyal, M., Chahal, S. and Kshatriya, G., (2008). Genetic affinities of the Siddis of South India: an emigrant population of East Africa. BMC Biol. *80*, 251-270.

4. Thangaraj, K., Ramana, G.V. and Singh, L., (1999). Y-chromosome and mitochondrial DNA polymorphisms in Indian populations. Electrophoresis. *20*, 1743-1747.

5. Ramana, G.V., Su, B., Jin, L., Singh, L., Wang, N., Underhill, P. and Chakraborty, R., (2001). Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. Eur J Hum Genet. *9*, 695-700.

6. Gauniyal, M., Aggarwal, A. and Kshatriya, G.K., (2011). Genomic Structure of the Immigrant Siddis of East Africa to Southern India: A Study of 20 Autosomal DNA Markers. Biochem Genet.

7. International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Bonnen, P.E., de Bakker, P.I.W. and Deloukas, P. et al., (2010). Integrating common and rare genetic variation in diverse human populations. Nature. *467*, 52-58.

8. Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M. and Waeber, G. et al., (2008). The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet. *83*, 347-358.

9. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. and Singh, L., (2009). Reconstructing Indian population history. Nature. *461*, 489-494.

10. Patterson, N., Price, A.L. and Reich, D., (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

11. Patterson, N., Petersen, D.C., van der Ross, R.E., Sudoyo, H., Glashoff, R.H., Marzuki, S., Reich, D. and Hayes, V.M., (2010). Genetic structure of a unique admixed population: implications for medical research. Hum Mol Genet. *19*, 411-419.

12. 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.

13. Reich, D.E. and Lander, E.S., (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502-510.

14. Scozzari, R., Torroni, A., Semino, O., Cruciani, F., Spedini, G. and Santachiara Benerecetti, S.A., (1994). Genetic studies in Cameroon: mitochondrial DNA polymorphisms in Bamileke. Hum Biol. *66*, 1-12.

15. Scozzari, R., Cruciani, F., Santolamazza, P., Sellitto, D., Cole, D.E., Rubin, L.A., Labuda, D., Marini, E., Succa, V. and Vona, G. et al., (1997). mtDNA and Y chromosome-specific polymorphisms in modern Ojibwa: implications about the origin of their gene pool. Am J Hum Genet. *60*, 241-244.

16. Scozzari, R., Cruciani, F., Pangrazio, A., Santolamazza, P., Vona, G., Moral, P., Latini, V., Varesi, L., Memmi, M.M. and Romano, V. et al., (2001). Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. Hum Immunol. *62*, 871-884.

17. Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., Destro-Bisol, G. and Coia, V. et al., (2002). A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet. *70*, 1197-1214.

18. Luis, J.R., Rowold, D.J., Regueiro, M., Caeiro, B., Cinnioglu, C., Roseman, C., Underhill, P.A., Cavalli-Sforza, L.L. and Herrera, R.J., (2004). The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. Am J Hum Genet. *74*, 532-544.

19. Semino, O., Santachiara-Benerecetti, A., Falaschi, F., Cavalli-Sforza, L. and Underhill, P., (2002). Ethiopians and Khoisan share the deepest clades of the human Y chromosome phylogeny. Am J Hum Genet. *70*, 265-268.

20. Semino, O., Magri, C., Benuzzi, G., Lin, A.A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P. and Oefner, P.J. et al., (2004). Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet. *74*, 1023-1034.

21. Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W., Kauffman, E., Bonne-tamir, B., Bertranpetit, J. and Francalacci, P. et al., (2000). Y chromosome sequence variation and the history of human populations. Nat Genet. *26*, 358-361.

22. Plaza, S., Salas, A., Calafell, F., Corte-Real, F., Bertranpetit, J., Carracedo, A. and Comas, D., (2004). Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Hum Genet. *115*, 439-447.

23. Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H.V. and Stepanov, V. et al., (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet. *72*, 313-332.

24. Thanseem, I., Thangaraj, K., Chaubey, G., Singh, V.K., Bhaskar, L.V.K.S., Reddy, B.M., Reddy, A.G. and Singh, L., (2006). Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. BMC Genet. *7*, 42.

25. Trivedi, R., Sahoo, S., Singh, A., Bindu, G., Banerjee, J., Tandon, M., Gaikwad, S., Rajkumar, R., Sitalaximi, T. and Ashma, R., (2008). Genetic Imprints of Pleistocene Origin of Indian Populations: A Comprehensive Phylogeographic Sketch of Indian Y-Chromosomes. Int J Hum Genet. *8*, 97-118.

26. Thangaraj, K., Naidu, B.P., Crivellaro, F., Tamang, R., Upadhyay, S., Sharma, V.K., Reddy, A.G., Walimbe, S.R., Chaubey, G. and Kivisild, T. et al., (2010). The influence of natural barriers in shaping the genetic structure of maharashtra populations. PloS one. *5*, e15283.

27. Brehm, A., Pereira, L., Bandelt, H.J., Prata, M.J. and Amorim, A., (2002). Mitochondrial portrait of the Cabo Verde archipelago: the Senegambian outpost of Atlantic slave trade. Ann Hum Genet. *66*, 49-60.

28. Berniell-Lee, G., Calafell, F., Bosch, E., Heyer, E., Sica, L., Mouguiama-Daouda, P., van der Veen, L., Hombert, J.M., Quintana-Murci, L. and Comas, D., (2009). Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. Mol Biol Evol. *26*, 1581-1589.

29. Tofanelli, S., Bertoncini, S., Castrì, L., Luiselli, D., Calafell, F., Donati, G. and Paoli, G., (2009). On the origins and admixture of Malagasy: new evidence from high-resolution analyses of paternal and maternal lineages. Mol Biol Evol. *26*, 2109-2124.

30. Zhivotovsky, L.A., Underhill, P.A., Cinnioglu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G. and Spedini, G. et al., (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet. *74*, 50-61.

31. Beerli, P. and Felsenstein, J., (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci U S A. *98*, 4563-4568.

32. Beerli, P. and Felsenstein, J., (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics. *152*, 763-773.

33. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M. and Howell, N., (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. *23*, 147.

34. van Oven, M. and Kayser, M., (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. *30*, E386-E394.

35. Pereira, L., Macaulay, V., Torroni, A., Scozzari, R., Prata, M.J. and Amorim, A., (2001). Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. Ann Hum Genet. *65*, 439-458.

36. Salas, A., Richards, M., De la Fe, T., Lareu, M.V., Sobrino, B., Sanchez-Diz, P., Macaulay, V. and Carracedo, A., (2002). The making of the African mtDNA landscape. Am J Hum Genet. *71*, 1082-111.

37. Salas, A., Richards, M., Lareu, M.V., Scozzari, R., Coppa, A., Torroni, A., Macaulay, V. and Carracedo, A., (2004). The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet. *74*, 454-65.

38. Beleza, S., Gusmão, L., Amorim, A., Carracedo, A. and Salas, A., (2005). The genetic legacy of western Bantu migrations. Hum Genet. *117*, 366-375.

39. Wood, E.T., Stover, D.A., Ehret, C., Destro-Bisol, G., Spedini, G., McLeod, H., Louie, L., Bamshad, M., Strassmann, B.I. and Soodyall, H. et al., (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet. *13*, 867-876.

40. Gonder, M.K., Mortensen, H.M., Reed, F.A., de Sousa, A. and Tishkoff, S.A., (2007). Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol. *24*, 757-768.

41. Crispim, D., Canani, L.H., Gross, J.L., Tschiedel, B., Souto, K.E.P. and Roisenberg, I., (2006). The European-specific mitochondrial cluster J/T could confer an increased risk of insulin-resistance and type 2 diabetes: an analysis of the m.4216T > C and m.4917A > G variants. Ann Hum Genet. *70*, 488-495.

42. Chaubey, G., Metspalu, M., Kivisild, T. and Villems, R., (2007). Peopling of South Asia: investigating the caste-tribe continuum in India. Bioessays. *29*, 91-100.

43. Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G. and 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.

Chapter 4

Reconstructing Roma history from genome-wide data*

*_Originally published as:_ Moorjani P, Patterson N, Loh PR, Lipson M, Kisfali P, Melegh BI, Bonin M, Kádaši L, Rieß O, Berger B, Reich D, Melegh B. (2013) Reconstructing Roma history from genome-wide data. *PLoS ONE*. 2013;8(3):e58633. doi: 10.1371/ journal.pone. 0058633. Epub 2013 Mar 13.

Author Contributions:

Conceived and designed the experiments: PM, NP, DR, BM

Performed experiments: BIM, PK, LK, BM

Analyzed the data: PM, NP, DR

Contributed data: MB, OR, DR

Contributed analysis tools: PM, PL, ML, NP, OR, DR

Wrote the paper: PM, PL, ML, DR, BM

4.1 Abstract

The Roma people, living throughout Europe and West Asia, are a diverse population linked by the Romani language and culture. Previous linguistic and genetic studies have suggested that the Roma migrated into Europe from South Asia about 1,000-1,500 years ago. Genetic inferences about Roma history have mostly focused on the Y chromosome and mitochondrial DNA. To explore what additional information can be learned from genome-wide data, we analyzed data from six Roma groups that we genotyped at hundreds of thousands of single nucleotide polymorphisms (SNPs). We estimate that the Roma harbor about 80% West Eurasian ancestry—derived from a combination of European and South Asian sources—and that the date of admixture of South Asian and European ancestry was about 850 years before present. We provide evidence for Eastern Europe being a major source of European ancestry, and North-west India being a major source of the South Asian ancestry in the Roma. By computing allele sharing as a measure of linkage disequilibrium, we estimate that the migration of Roma out of the Indian subcontinent was accompanied by a severe founder event, which appears to have been followed by a major demographic expansion after the arrival in Europe.

4.2 Introduction

The Roma (also called Romani) are a unique and diverse population that live in Europe, Near East, Caucasus, and the Americas. They speak more than 60 dialects of a rapidly evolving language called *Romani* and belong to various social and religious groups across Europe. Their census size has been estimated to be in the range of 10-15 million[1], with the largest populations in Eastern Europe[2]. They do not have written history or genealogy (as Romani does not have a single convention for writing) and thus most of the information about their history has been

inferred based on linguistics, genetics and historical records of the countries where they have resided.

Historical studies have suggested that the Roma are originally from India, and that they migrated to Europe between the 5th and 10th century[3]. It has been argued that their migration route included Persia, Armenia, Anatolia, and Greece[3; 4]. The Roma then settled in multiple locations within Europe and were widespread in Europe by the 15th century; descendents of these migrants currently live primarily in the Balkans, Spain, and Portugal[5].

Anthropological and linguistic studies have documented striking similarities between the cultures and languages of various Indian groups and Roma. Social structure in Roma groups is similar to the *castes* of India, where the groups are often defined by profession[2; 3]. Like many Indian populations, the Roma practice endogamy and individuals of one Roma clan (sub-ethnic group) preferentially marry within the same group, and marriages across clans are proscribed[3]. Anthropological studies have also suggested a link between the Roma and Banjara (nomadic gypsy groups) residing in India[3] (even though linguistic analysis of the *Banjari* or *Lamani*, languages spoken by the Indian nomadic groups, have little similarity to Romani[6]). Comparative linguistics have further suggested that Northwestern Indian languages like Punjabi or Kashmiri or Central Indian languages like Hindi are most closely related to Romani[7; 8].

Genetics provides a complementary source of information to data from history, archaeology and linguistics. Y-chromosome marker H1a-M82 and mitochondrial haplogroups M5a1, M18 and M35b that are thought to be characteristic of South Asian ancestry, are present at high frequency in Roma populations[9; 10]. However, there is no consensus about the specific ancestral group/ geographic region within South Asia that is most closely related to the ancestral population of the Roma. A recent study based on Y-chromosome markers showed that the Roma

descended from southern Indian groups[11], which contradicts previous reports based on mtDNA haplogroups that have placed the origin of Roma in Northwest India. While mtDNA and Y chromosome analyses provide valuable information about the maternal and paternal lineages, a limitation of these studies is that they represent only one instantiation of the genealogical process. Autosomal data permits simultaneous analysis of multiple lineages, which can provide novel information about population history.

Here we analyze whole genome SNP array data from 27 Roma samples belonging to six groups sampled from 4 countries in Europe (three separate ethnic groups from Hungary, and one group each from Romania, Spain and Slovakia). Our aim was to address the following questions: (1) What is the source of the European ancestry in the Roma? (2) What is the relationship of the Roma to the present-day South Asian populations? (3) What is the proportion and timing of major gene flow into this population? (4) Can we characterize the founder events that have occurred in the history of this population?

4.3 Results

Genome-wide ancestry analysis of the Roma

We applied Principal Component Analysis (PCA) using the SMARTPCA software[12] and the clustering algorithm ADMIXTURE[13] to study the relationship of Roma to other worldwide populations in a merged dataset of Roma and HapMap populations. In PCA, the Roma fall between the South Asians (Gujaratis) and Europeans, consistent with Roma deriving ancestry both South Asians and Europeans and in line with previous mtDNA and Y chromosome analyses[9; 10] (Figure 4.1). The ADMIXTURE software, which implements a maximum likelihood method to infer the genetic ancestry of each individual modeled as a mixture of K ancestral groups, produces very similar inferences[13]. At K=6 (which has the lowest cross-validation error),

we observe clustering based on major continental ancestry. Similar to the PCA results, the Roma individuals cluster with South Asians and Europeans (Figure 4.1, Figure C.1). We also examined pairwise average allele frequency differentiation (F_{st}) between Roma and major continental groups (see Table C.1) and observed that they have the lowest F_{st} with other European groups.

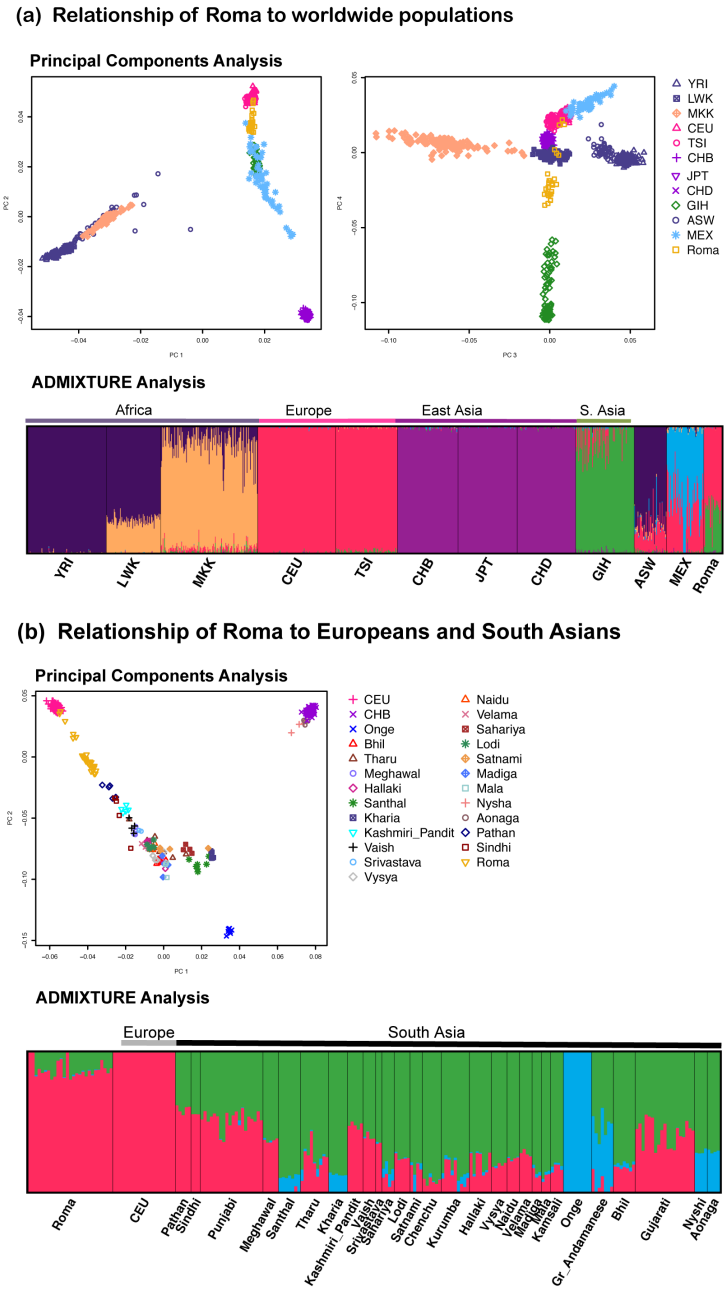


Figure 4.1 Relationship of Roma with other worldwide populations

Figure 4.1 (Continued). We applied PCA and ADMIXTURE to study the relationship of Roma with the HapMap and South Asian populations. In PCA, each point represents an individual, and in ADMIXTURE, each line represents an individual. (a) shows the PCA and ADMIXTURE results for clustering of Roma and HapMap populations. The populations codes are as follows: Yoruba in Ibadan, Nigeria (YRI), Luhya in Webuye, Kenya (LWK), Maasai in Kinyawa, Kenya (MKK), Utah residents with Northern and Western European ancestry (CEU), Toscani in Italia (TSI), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT), Chinese in Metropolitan Denver, Colorado (CHD), Gujarati Indians in Houston, Texas (GIH), African ancestry in Southwest USA (ASW) and Mexican ancestry in Los Angeles, California (MEX), and (b) shows the PCA and ADMIXTURE results for clustering of Roma and South Asian groups. We limit the sample size of all groups (except Roma) to 20 individuals.

Previous studies have shown that the HapMap Gujarati population is not an ideal surrogate for the variation in India, as this group is heterogeneous and has recent West Eurasian ancestry[14]. To study the relationship of Roma to South Asians, we repeated the clustering analysis with Roma, Europeans and 28 South Asian groups (24 Indian groups from the India Project (we remove Siddis as they have recent African ancestry), Pathan and Sindhi from HGDP and Punjabi and Gujarati from POPRES). As previously seen in PCA, we observed that all Indians fall on a cline of variable relatedness to Europeans and indigenous Andamanese population (Onge)[14]. The Roma also fall on this cline but they appear to be closest to the European cluster compared to any other South Asian group included (Figure 4.1b). Similar results were observed in our ADMIXTURE analysis (Figure 4.1b, Figure C.1). Based on the PCA and ADMIXTURE analysis, we excluded three Roma outlier samples from further analyses, as they appeared to have very recent admixture from neighboring non-Roma European populations (likely in the past few generations).

We applied the *4 Population Test*[14] to formally examine if the Roma have evidence of a mixture of European and South Asian ancestry. We used individuals of Northern European ancestry (CEU) and Andamanese (Onge) as surrogates for the European and South Asian ancestral populations respectively. We tested whether the phylogenetic tree (Africans, Europeans, South Asians, Roma) is consistent with the data. We choose Onge for this analysis, since, unlike their distant relatives on the Indian mainland, they do not have any evidence of West Eurasian related admixture[14]. Applying the *4 Population Test*, we observed highly significant violations of the expected phylogenetic tree topology, confirming that the Roma are admixed; that is, they have ancestry from both South Asians and Europeans (Table C.2). We note that this test does not distinguish between European and West Asian ancestry and qualitatively similar results would be observed if we replace CEU with any other West Eurasian population (other groups from Europe, Middle East, Central Asia or Caucasus), hence we refer to this ancestry component as "Ancestral West Eurasian (AWE)".

To quantify the magnitude of the South Asian and West Eurasian ancestry in the Roma, we applied F_4 *Ratio Estimation*[15] using the model shown in Figure C.2, which can estimate admixture proportions in the absence of data from good surrogates of the ancestral populations. Here, we used CEU and Adygei (a population from the Caucasus) represent the West Eurasian component and Onge to represent the ancestral South Asian component (referred to as Ancestral South Indian (ASI)) as they do not have any West Eurasian ancestry [14]. The F_4 *Ratio Estimation* is known to work only if we have access to data from populations that form a clade with the unadmixed ancestral populations. Since all populations in mainland India are admixed none are appropriate for this test[14]. To further evaluate our model of population relationships in Figure C.2, we used *admixture graph*[15] and found that this model provides a good fit to the data.

Applying the *F_4 Ratio Estimation* to Roma (pooling all samples together), we estimate that the Roma have on average 77.5 ± 1.8% West Eurasian related ancestry (standard errors were computed using a Block Jackknife with a block size of 5cM) (Table C.2). As all Indian groups harbor ancestry from a West Eurasian related populations (previously referred to as Ancestral North Indian (ANI) ancestry[14]), we note that some of West Eurasian related ancestry we detect in Roma likely derives from India itself—from the ANI—while other parts may be from European or Middle Eastern admixture (post exodus from India).

Estimating a date of European admixture in the Roma

To infer the date of the gene flow, we applied a modified version of *ROLLOFF*[16], which uses the decay of admixture linkage disequilibrium (LD) to estimate the time of admixture. *ROLLOFF* computes SNP correlations in the admixed population and weights the correlations by the allele frequency difference in the ancestral populations such that the signal is sensitive to admixture LD. While this method estimates accurate dates of admixture in most cases, we observed that it is noticeably biased in case of strong founder events post admixture (Table C.3). The bias is related to a normalization term that exhibits an exponential decay behavior in the presence of a strong founder event, thus confounding the admixture date (see details in Note C.1, Figure C.3). We propose a modification to the *ROLLOFF* statistic that removes the bias (Note C.1, Table C.3). In addition, the new statistic computes covariance instead of correlation between SNPs; this does not affect the performance of the method but makes it mathematically more tractable. Throughout the manuscript, we use the modified *ROLLOFF* statistic ($R(d)$) unless specified otherwise. Simulations show that this statistic gives accurate and unbiased results up to 300 generations (Note C.2, Figure C.4).

A feature of *ROLLOFF* is that it uses allele frequency information in the ancestral populations to amplify the admixture signal relative to background LD. While data from the ancestral populations is not available for Roma, this information can be obtained by performing PCA using present day Europeans and South Asians. Simulations show that using PCA-based SNP loading effectively captures the allele frequency differentiation between the ancestral populations and can be used for estimating dates of mixture (Note C.2, Figure C.5).

Applying the *ROLLOFF* (using *R(d)*) to the Roma samples with the SNP loading estimated using PCA of Europeans (CEU) and 16 Indian groups (limited to groups that fall on the main cline of West Eurasian relatedness in PCA so that the signal is not confounded by other ancestry components), we estimate that the West Eurasian admixture in Roma occurred 29 ± 2 generations or about 780-900 years ago, assuming one generation = 29 years[17] (Figure 4.2). This is consistent with mixture having occurred only after the historically recorded arrival of the Roma in Europe between 1,000-1,500 years ago[3].

A potential complication is that the date we are estimating may also be reflecting earlier admixture with ANI in India and any gene flow from Middle Eastern populations that occurred after the Roma exodus from India. The allele frequency of ANI and Middle Eastern populations are correlated to the allele frequencies of the Europeans used in the analysis, and hence the date of admixture inferred using a single exponential function should be interpreted as an average date of all West Eurasian related gene flow events. When we consider a two-pulse model of admixture (by fitting a sum of two exponential functions to infer the dates), we obtain dates of 37 and 4 generations. The older date corresponds to about 1,000 years before present – again consistent with the historical record – and both dates are much more recent than any estimates

obtained by applying *ROLLOFF* in India. This suggests that the admixture we are detecting is genuinely related to events that occurred after the exodus from India.

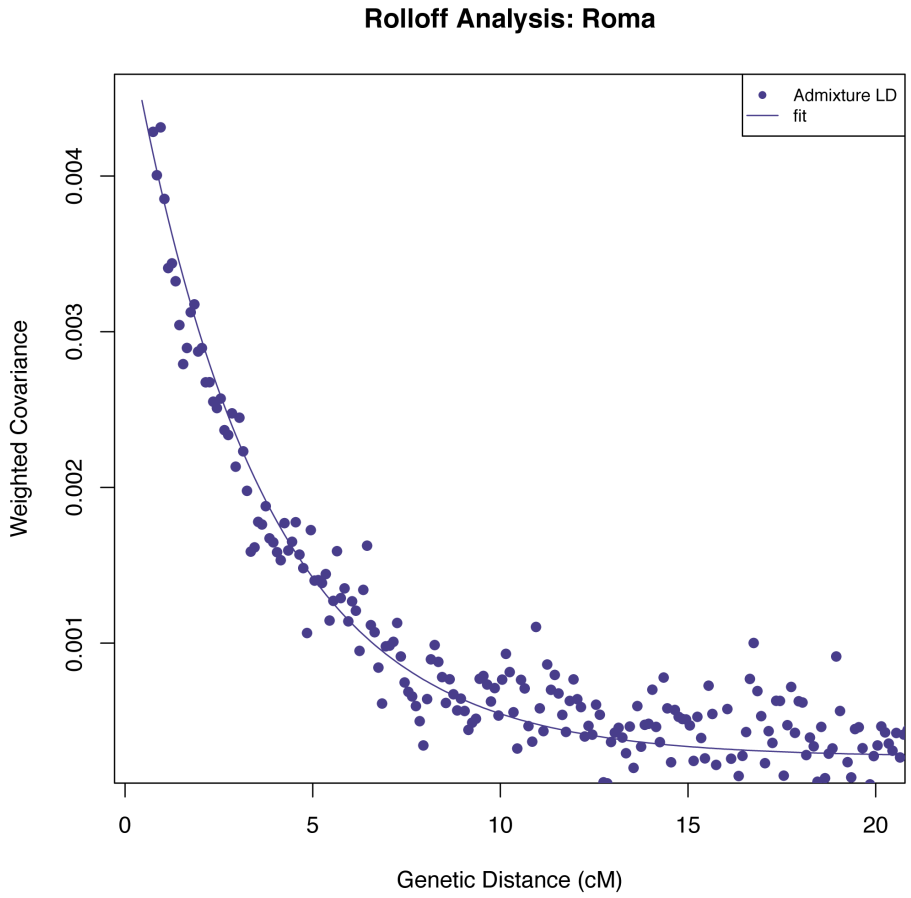


Figure 4.2 Admixture date estimation. We performed *ROLLOFF* (using *R(d)*) on the Roma samples (n = 24). We plot the weighted covariance as a function of genetic distance, and obtain a date by fitting an exponential function with an affine term: $y = Ae^{-nd} + c$, where d is the genetic distance in Morgans and n is the number of generations since mixture. We do not show inter-SNP intervals of <0.5cM since we have found that at this distance admixture LD begins to be confounded by background LD.

Source of the European ancestry in Roma

To learn about the relationship of the Roma to European populations, we estimated the pairwise Identity-by-descent (IBD) sharing between each Roma individual and non-Roma European individual. We grouped the European samples from POPRES, HapMap and HGDP into four major regional groups: Northern (n = 595), Southern (n = 649), Eastern (n = 82), and Western Europe (n = 241). IBD segments (>3 centimorgans (cM)) were detected using GERMLINE[18]. Next, we computed an average pairwise sharing distance between Roma and the European groups in each region (see Methods). We observed that Roma exhibit the highest IBD sharing with individuals from Eastern Europe (Figure 4.3a). When we perform stratified analysis (where Roma individuals from each country were considered separately), we observed that the highest sharing for each Roma group is still with Eastern Europeans (even for Roma individuals from Spain) (Figure C.6).

Source of the South Asian ancestry in Roma

To learn about the source of the South Asian ancestry in Roma, we inferred the pairwise IBD sharing distance between Roma and various South Asian groups. Again, we performed GERMLINE analysis to compute the average pairwise sharing distance between Roma and 28 South Asian populations (from India Project, HGDP and POPRES). To simplify the analysis, we classified the samples into 8 groups based on geographical region within India: North (n = 38), Northwest (n = 225), Northeast (n = 8), Southwest (n = 16), Southeast (n = 29), East (n = 11), West (n = 32), and Andamanese (n = 16). We observe that the Roma share the highest proportion of IBD segments with groups from the Northwest of India (Figure 4.3b). Interestingly, the two Northwest Indian groups that show the highest relatedness to Roma (Punjabi, Kashmiri Pandit)

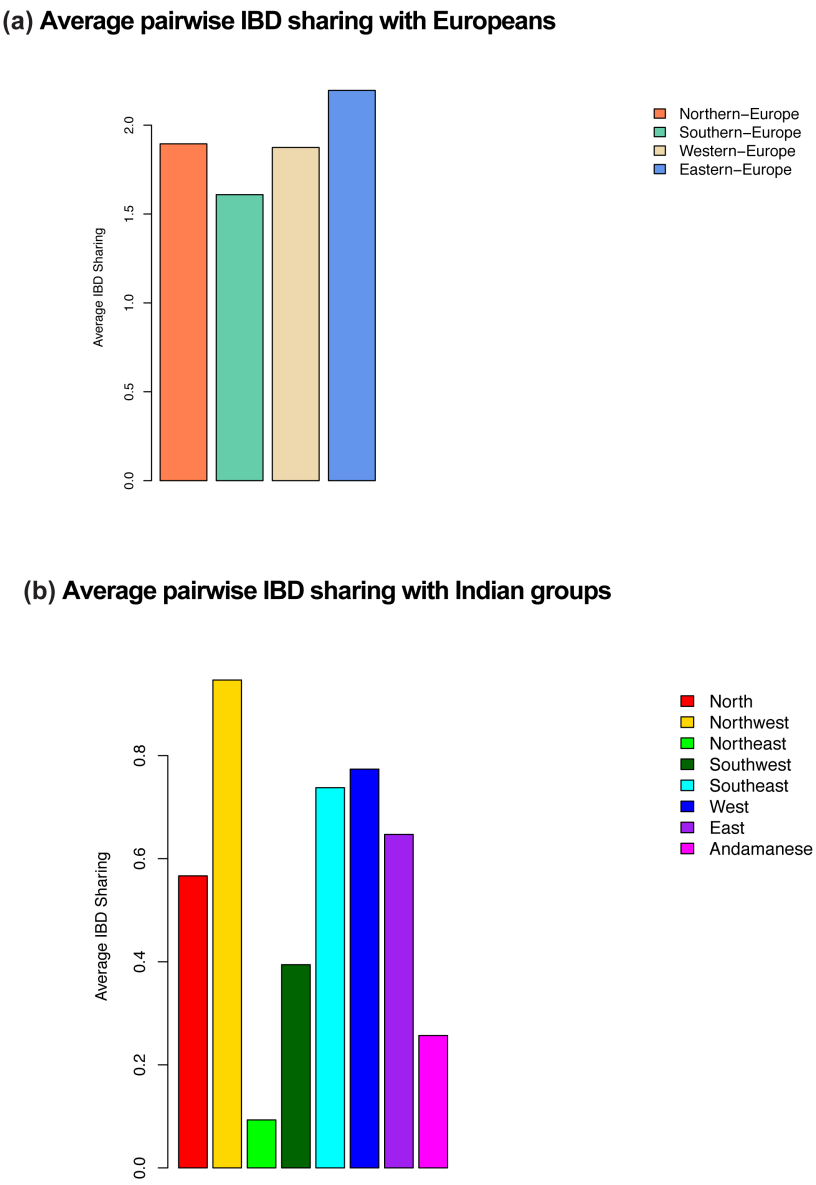


Figure 4.3 The European and South Asian sources of Roma ancestry. We computed a genome-wide average IBD sharing distance between Roma (all samples combined in one group) and other regional groups. Details of the regional grouping are described in Methods. (a) shows the average pairwise IBD sharing between Roma and Europeans (grouped into four regional categories), (b) shows IBD sharing average pairwise IBD sharing between Roma and South Asians (grouped into 8 regional categories).

are also the populations that have highest proportion of West Eurasian-related (ANI) ancestry in our sample. To control for the possibility that the high IBD sharing could be an artifact related to high ANI ancestry, we recalculated the IBD sharing regressing out the ANI ancestry proportion and observed that the Roma continue to share the highest IBD segments with the Northwest Indian groups (Note C.3). These findings are consistent with analyses of mtDNA that also place the most likely South Asian source of the Roma in Northwest India[10].

An important caveat is that we have large variation in the number of samples from each regional group, with some groups containing only a handful of samples. In order to control for the sample sizes, we performed bootstrap analysis drawing a random sample of up to 30 individuals from each regional group and recomputed the IBD statistics. We repeated the process 100 times and estimated the mean and standard error (Note C.3). We observed that Roma continue to share the highest IBD segments with Northwest Indian groups. There is very little variability across the 100 runs, suggesting that this analysis may also be picking up founder events shared between Roma and Indian groups (Note C.3, Figure C.7).

Characterizing the founder events

Previous genetic and social studies have shown that the present day Roma population has descended from a small number of ancestors with subsequent genetic and cultural isolation[10; 19]. A history of founder events in a population can lead to an increase in homozygosity and large stretches of allele sharing across individuals within the same population. This can be measured by estimating the proportion of the autosomal genome that has homozygous genotypes. We applied PLINK v1.07[20] to compute a genomic measure of individual autozygosity for all Roma individuals and 30 random individuals from each of the 11 HapMap populations. PLINK uses a

sliding window approach to find regions of the genome that are at least 1MB in length and contain 100 contiguous homozygous SNPs. For each individual, we computed the number and overall length of the autozygous segments and observed that the Roma have very high levels of autozygosity compared to other HapMap populations (Figure 4.4a). This suggests that inbreeding (or consanguineous marriages) might be common in Roma.

To infer the date of the founder event in Roma, we studied the relationship of allele sharing with increasing distance as reported in Reich et al (2009)[21]. This statistic is based on examining the autocorrelation of allele sharing between pairs of individuals within a population, and then subtracting the cross-population autocorrelation to remove the effects of ancestral allele sharing inherited from the common ancestor. By measuring the exponential decay of auto-correlation with genetic distance, we obtained an estimate of the age of the founder event. Simulations have shown that this method can accurately estimate the dates of recent founder events, even in admixed populations (Note C.4, Table C.5).

Applying this method to Roma and subtracting the shared Roma and European (CEU) autocorrelation, we estimate that a Roma founder event occurred 27 generations or ~800 years ago (assuming one generation = 29 years[17]) (Figure 4.4b). This is consistent with reports that the Roma exodus from India occurred 1,000 years ago[3], and suggests that the migration out of the Indian sub-continent may have been associated with a significant founder event in which a small number of ancestral individuals gave rise to the present-day Roma population.

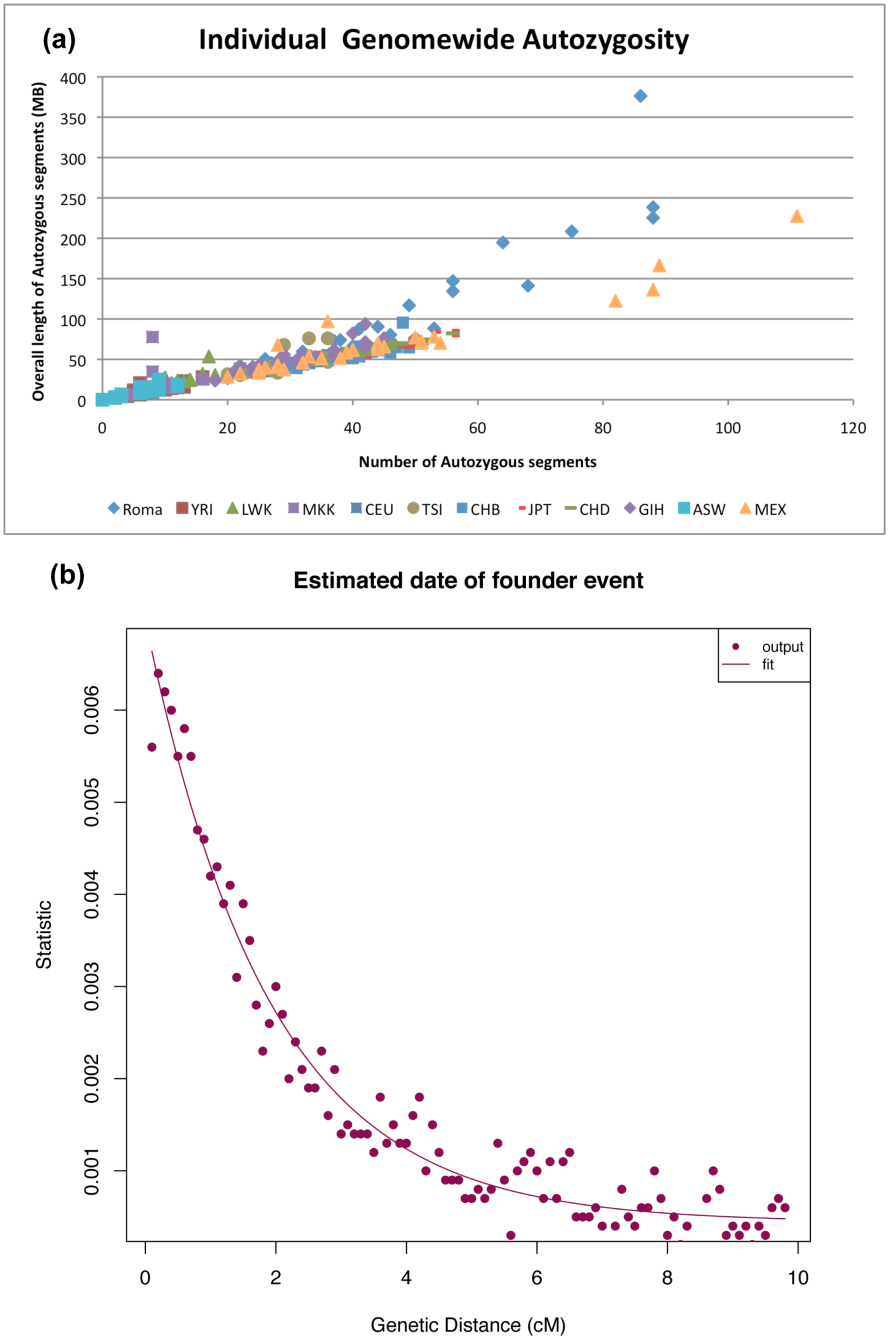


Figure 4.4 Founder events in the Roma. (a) shows estimates of genomewide autozygosity in Roma and individuals from HapMap (n = 30 from each of the 11 HapMap populations). Each point represents an individual with the color-coding described in the legend. **(b)** shows the decay of autocorrelation with genetic distance. We fit an exponential function: $y = Ae^{-2tD} + c$ where D = distance in Morgans and t = time of founder event. We thus infer a founder event date of 27 generations.

4.4 Discussion

Using genome-wide SNP data from Roma individuals, we have provided (1) confirmation of previous mtDNA and Y chromosome results with autosomal data, and (2) some new insights that take special advantage of autosomal data.

We have performed formal tests to confirm that Roma are admixed and have ancestry from two highly divergent populations: a West Eurasian population and a South Asian population. We estimate that the Roma have 77.5% West Eurasian ancestry, reflecting a combined estimate of the ANI ancestry that the Roma derive from their South Asian ancestors (pre-exodus) and the European ancestry related to the admixture in Europe (post-exodus from India). Our estimate of West Eurasian ancestry is broadly consistent with admixture proportions estimated using autosomal short tandem repeats (66-100%)[22]. Our estimates of non-West Eurasian ancestry (ASI = 22.5 ± 1.8%) are also consistent with the estimates from mitochondrial DNA (26.5%) and Y-chromosome (16.7%) markers[23; 24].

Our identity-by-descent analysis provides novel insights related to the source of the ancestral populations of Roma. We provide evidence for Eastern Europe being a major source of the European ancestry, and Northwest India being a major source of the South Asian ancestry in Roma. Our inferences about the geographic origin within South Asia help resolve a long-standing debate related to the origin of the Romani people. Our results are consistent with reports from linguistics[7] and mtDNA studies[10], which have shown that present day Northwest Indian populations (from Kashmir and Punjab), are candidates for being the source of the Indian ancestry in Roma[10; 23]. However, we caution that IBD based methods require large sample sizes to be well powered to detect subtle differences between geographic regions.

A historically informative insight from our analysis is the date of the West Eurasian gene flow into Roma. Using a statistic that captures the pattern of admixture related linkage disequilibrium; we estimate that the admixture between Roma and West Eurasians occurred 29 ± 2 generations or about 780-900 years ago. The earliest records of the arrival of Roma in the Balkans dates back to the 11th-12th century[3], which is concordant with our estimated date of mixture[3]. It is important to note that the Roma have ancestry from both ANI and Europeans and thus the estimated date of admixture with Europeans (post exodus) is slightly downward biased (older). Simulations have shown in the case of two gene flow events, the date of admixture estimated by *ROLLOFF* tends to reflect the date of the more recent gene flow event as the interval between the dates of two gene flow events increases (Table C.4, Note C.2).

Disease mutation screening in the Roma has shown that they have an increased proportion of private mutations[19]. For example, deletion 1267delG is known to cause a neuromuscular disorder, *congenital myasthenia*, and has a high carrier frequency in many Roma groups that reside in different parts of Europe. This mutation has only been observed in South Asian populations previously[19; 25]. This provides evidence that the different Roma groups have a history of a shared founder events with South Asians. In order to obtain temporal information of the founder event that has likely increased the frequency of such disease causing mutations in Roma, we studied LD based allele sharing statistics and estimated that the founder event in Roma occurred about 27 generations, or 800 years, ago. This agrees with previous reports from Morar et al. (2004)[25] who hypothesize that the entire Roma population was founded about 32-40 generations ago.

After this manuscript was submitted, two other studies characterizing the population history of Roma were published. First, a study based on Y-chromosome haplogroups showed

that on the paternal lineage, Roma haplotypes cluster predominantly with the Northwestern Indian haplotypes[26], consistent with our findings based on autosomal IBD sharing. The second study was based on whole genome SNP genotype data like ours [27]. Our findings are broadly consistent with the results from that paper, although with some notable differences. For inferring the date of the founder event, the other study uses a two-pulse model (an out-of-India founder event, followed by a second founder event that affects only the western Roma groups). We instead estimate the date of a single shared founder event; with our limited sample size (we have only 2 samples from western Roma groups), we cannot recover the entire distribution of founder events and so the date of the founder event in our study should be interpreted as an average date of multiple founder events. Similarly, the other study, using a continuous admixture model, estimates that the admixture in Roma occurred over a period of 38 generations[27]. Assuming a single admixture model, we estimate that the average date of admixture is 29 ± 2 generations. However, when we consider a two-pulse model of admixture, we infer the dates of 37 and 4 generations, consistent with the results of the other study.

In conclusion, our study has confirmed that the Roma have ancestry from South Asians (likely Northwest Indians) and West Eurasians (likely Eastern Europeans), with mixture occurring around 30 generations ago and major founder events shortly afterward. An important opportunity for future work is to perform homozygosity mapping in Roma that can aid in finding disease-causing mutations related to the founder events.

4.5 Materials and Methods

Datasets

We collected 27 Roma samples belonging to six groups that were sampled from four countries in Europe from Hungary (3 linguistically and culturally separated sub-groups: 7 samples from Olah (Vlah), 4 samples from Beas (Boyash) and 4 samples from Romungro), 4 samples from Romania, 4 samples from Spain and 4 samples from Slovakia (Slovakian speaking Roma). All research involving human participants was approved by the Regional Ethics Committee Board (REKEB) and the Hungarian National Ethics Committee (ETT TUKEB). Each study participant attended a 45-60mins verbal orientation session about the study design and goals and then provided written informed consent. All the research was conducted according to the principles expressed in the Declaration of Helsinki. Roma individuals self-reported as being descendants of the same tribe for at least three generations. The samples were genotyped using an Affymetrix 1M SNP chip. We required <5% missing genotypes per sample per SNP to be included in the analysis (27 individuals, 726,404 SNPs passed this threshold). These data were merged with data from four other sources, including the International Haplotype Map Phase 3 (HapMap) (n = 1,115 samples from 11 populations genotyped on Affymetrix 1M array)[28], the CEPH-Human Genome Diversity Panel (HGDP) (n = 257 individuals from 51 populations genotyped on Affymetrix 500K SNP array)[29; 30], our previous study of Indian genetic variation which we call the "India Project" in this paper (n = 132 individuals from 25 groups genotyped on an Affymetrix 1M SNP array)[14] and the Population Reference Sample (POPRES) (n = 3,845 individuals from 37 European populations genotyped on an Affymetrix 500K SNP array)[31]. Depending on the analyses, we included different number of reference populations from these sources.

Population Structure Analysis and F_{st} calculation

To study the relationship of Roma with HapMap populations, we created a merged dataset of Roma and HapMap populations (n = 1,142 and 726,404 SNPs). As background LD can affect both PCA and ADMIXTURE analysis, we thinned the marker set using PLINK v1.07 [20] by excluding SNPs in strong LD (pairwise genotypic correlation $r^2 > 0.1$) in a window of 50 SNPs (sliding the window by 5 SNPs at a time). The thinned dataset contained 61,052 SNPs. We used SMARTPCA[12] to perform PCA and to compute F_{ST} values. Clustering analysis was performed using ADMIXTURE[13].

To study the relationship of Roma with South Asians, we created a merged dataset of Roma, HapMap, POPRES and HGDP (n = 1,966 and 205,710 SNPs) and performed PCA and ADMIXTURE using the LD thinned dataset containing 55,303 SNPs.

Formal tests of population mixture

To test if Roma have West Eurasian and Indian ancestry, we used the unrooted phylogenetic tree ((YRI, CEU), (Onge, Roma)) and computed the *4-population test* statistic for all three phylogenetic trees that can possibly relate these populations. For this analysis, we created a merged dataset of Roma, India project and HapMap populations (n = 1,274 and 524,053 SNPs). Let YRI_i, CEU_i, $Onge_i$ and $Roma_i$ be the allele frequencies for SNP i in the populations YRI, CEU, Onge and Roma respectively. Specifically, we compute the correlation: $\rho(YRI_i\text{-}CEU_i, Onge_i\text{-}Roma_i)$ for all SNPs across the genome. In the absence of mixture, the expected correlation would be 0. Standard errors were computed using Block Jackknife[32; 33] where a block of 5cM was dropped in each run.

Estimating genome-wide ancestry proportion

We estimate the genome-wide proportion of ancestry using *F_4 Ratio Estimation*[15] which estimates the excess of West Eurasian ancestry compared to Onge. We use the model of population relationships shown in the Figure C.2. We test this model using *admixture graph*[15] and find that the model is a good fit to the data (meaning that none of the f-statistics are greater than three standard errors from expectation). *F_4 Ratio Estimation* computes the ratio of f_4(YRI$_i$, Adygei$_i$; Roma$_i$-Onge$_i$)/ f_4(YRI$_i$, Adygei$_i$; CEU$_i$-Onge$_i$). This quantity is summed over all sites (262,558 SNPs) and the standard errors are computed using the Block Jackknife (block size of 5cM). To represent all the populations needed for this analysis, we created a merged dataset that included data from the Roma, the India project, HGDP and HapMap (n = 1,531 and 262,558 SNPs).

GERMLINE analysis

IBD segments were detected using GERMLINE[18]. For this analysis, we phased the data from all relevant populations using Beagle[34] and then ran GERMLINE in genotype extension mode on a combined dataset of Roma, HapMap, India Project, POPRES and HGDP (n = 1,966 and 205,710 SNPs). We applied the following parameters for calculating IBD segments: seed size = 75, minimum IBD segments length = 3cM, and the number of heterozygous or homozygous errors = 0. The output of GERMLINE was used to compute an average pairwise sharing between populations I and J as previously reported in reference[35].

$$\text{Average sharing} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} IBD_{ij}}{n \times m}$$

where IBD_{ij} = the length of IBD segment shared between individual *i* and *j* and *n*, *m* are the number of individuals in population *I* and *J* respectively.

For identifying the source of the European ancestry, we computed the average sharing between Roma and each of the four geographic regions in Europe. Each group contained the following samples: *Northern-Europe* (*n* = 595) included CEU from HapMap, Orcadian from HGDP, and Latvia, United Kingdom, Ireland, Sweden, Scotland, Norway, Denmark, and Finland from POPRES, *Southern-Europe* (*n* = 649) included TSI from HapMap, Italian, Basque, Sardinian, and Tuscan from HGDP, and Spain, Croatia, Bosnia-Herzegovina, Albania, Macedonia, Slovenia, Kosovo, Italy, Cyprus, Portugal, Greece, and Serbia from POPRES, *Eastern-Europe* (*n* = 82) included Russian from HGDP and Romania, Hungary, Slovakia, Czech Republic, Bulgaria, Ukraine, Poland, and Russia from POPRES, and *Western-Europe* (*n* = 241) included French from HGDP and Germany, Belgium, France, Austria, and Netherlands from POPRES.

Similarly, for identifying the source of the South Asian ancestry we computed average IBD distance between Roma and South Asians. We grouped the South Asian samples in seven regional categories as follows: *North* (*n* = 38) included Tharu, Kharia, Vaish, Srivastava, Sahariya, Lodi, HGDP Pathan and Sindhi, *Northwest* (*n* = 225) included Kashmiri Pandit and POPRES Punjabi, *Northeast* (*n* = 8) included Nyshi and Ao Naga, *Southwest* (*n* = 16) included Kurumba and Hallaki, *Southeast* (*n* = 29) includes Madiga, Mala, Vysya, Chenchu, Naidu, Velama and Kamsali, *West* (*n* = 32) included Bhil, Meghawal and POPRES Gujarati, *East* included Santhal and Satnami, and *Andamanese* (*n* = 16) included Great Andamanese and Onge.

Estimation of a date of mixture

We applied modified *ROLLOFF*[16] to estimate the date of mixture in a combined dataset containing 1,274 individuals and 524,053 SNPs. For each pair of SNPs (*x,y*) separated by a distance *d* Morgans, we compute covariance between (*x,y*). Specifically, we use the following statistic -

$$R(d) = \frac{\sum_{|x-y| \approx d} z(x,y) w(x,y)}{\sum_{|x-y| \approx d} w(x,y)^2}$$

where $z(x,y)$= covariance between SNPs (*x, y*) and weight function $w(x,y)$ = a weight function that can be the allele frequency difference between the ancestral populations or the PCA based loadings for SNPs (*x, y*). We study the relationship of the weighted covariance with genetic distance, and obtain a date by fitting an exponential function with an affine term $y = Ae^{-nd} + c$, where *n* is the number of generations since admixture, *d* is the distance in Morgans, *c* is the affine term (non-zero asymptote of the fitted curve) and *A* is amplitude of the weighted LD curve (LD at short distances). Standard errors were computed using a weighted Block Jackknife[32; 33] where one chromosome was dropped in each run. We fit a sum of exponentials to estimate the dates of admixture under a two-pulse model of admixture using the exponential function: $y = Ae^{-n_1 d} + Be^{-n_2 d} + c$, where n_1 , n_2 are the admixture dates in generations.

Estimating individual autozygosity

We used PLINK v1.07[20] to identify autozygous segments in the genome in a combined dataset of 1,274 individuals and 524,053 SNPs. PLINK uses a sliding window approach to find regions of the genome that are at least 1MB in length and contains 100 contiguous homozygous SNPs. We allowed one heterozygous and five missing calls per segment. Autozygous segments were identified separately for each individual. We applied this method to compute genomic

autozygosity (overall length of autozygous segments) for each Roma and 30 random individuals from each HapMap population.

Estimating a date for the founder event

To infer the date of the founder event, we compute the correlation of allele sharing as a measure of LD as described in reference[14] using a dataset containing Roma and HapMap populations (n = 1,142 and 726,404 SNPs). Specifically, we compute the autocorrelation of allele sharing between pairs Roma individuals, and then subtract the (Roma, CEU) cross-population autocorrelation to remove the effects of ancestral allele sharing. We thus get a measure for the Roma-specific LD related to the excess of allele sharing in this group. We plot the auto-correlation against genetic distance to infer the time of founder event. Specifically, we fit the exponential function: $y = Ae^{-2tD} + c$, where D = distance in Morgans and t = time of founder event.

Acknowledgements

We thank Kasia Bryc and Sriram Sankararaman for helpful discussions and constructive comments on the manuscript. Detailed information about the methods and sample collection for the POPRES data are described in Nelson et al. (2008). The dataset was obtained from dbGAP (accession number: phs000145.v1.p1).

4.6 References

1. Liégeois, J.P. (1994). Roma, gypsies, travellers.(Sales Agent Manhattan Pub. Co. Distributor).

2. Marushiakova, E., and Popov, V. (1997). Gypsies (Roma) in Bulgaria.(P. Lang).

3. Fraser, A.M. (1995). The gypsies.(Wiley-Blackwell).

4. Kalaydjieva, L., Morar, B., Chaix, R., and Tang, H. (2005). A newly discovered founder population: the Roma/Gypsies. Bioessays 27, 1084-1094.

5. Schurr, T.G. (2004). Reconstructing the origins and migrations of diasporic populations: the case of the European Gypsies. American anthropologist 106, 267-281.

6. Trail, R.L. (1970). The grammar of Lamani.(Summer Institute of Linguistics of the University of Oklahoma).

7. Boerger, B.H. (1984). Proto-Romanes phonology. Dissertation.

8. Turner, R.L. (1927). The Position of Romani in Indo-Aryan. Gypsy Lore Society. In. (Monographs.

9. Pamjav, H., Zalán, A., Béres, J., Nagy, M., and Chang, Y.M. (2011). Genetic structure of the paternal lineage of the Roma People. American Journal of Physical Anthropology 145, 21-29.

10. Mendizabal, I., Valente, C., Gusmão, A., Alves, C., Gomes, V., Goios, A., Parson, W., Calafell, F., Alvarez, L., and Amorim, A. (2011). Reconstructing the Indian Origin and Dispersal of the European Roma: A Maternal Genetic Perspective. PloS one 6, e15988.

11. Regueiro, M., Rivera, L., Chennakrishnaiah, S., Popovic, B., Andjus, S., Milasin, J., and Herrera, R.J. (2012). Ancestral modal Y-STR haplotype shared among Romani and South Indian populations. Gene 504, 296-302.

12. Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2, e190.

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

14. Reich, D., Thangaraj, K., Patterson, N., Price, A., and Singh, L. (2009). Reconstructing Indian population history. Nature 461, 489-494.

15. Patterson, N.J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. Genetics.

16. 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 Genetics 7, e1001373.

17. Fenner, J. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology 128, 415.

18. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Research 19, 318.

19. Kalaydjieva, L., Gresham, D., and Calafell, F. (2001). Genetic studies of the Roma (Gypsies): a review. BMC Medical Genetics 2, 5.

20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., and Daly, M. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559-575.

21. Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., and Johnson, P.L.F. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053-1060.

22. Gusmão, A., Valente, C., Gomes, V., Alves, C., Amorim, A., Prata, M.J., and Gusmão, L. A genetic historical sketch of European Gypsies: The perspective from autosomal markers. American Journal of Physical Anthropology 141, 507-514.

23. Gusmão, A., Gusmão, L., Gomes, V., Alves, C., Calafell, F., Amorim, A., and Prata, M. (2008). A Perspective on the History of the Iberian Gypsies Provided by Phylogeographic Analysis of Y-Chromosome Lineages. Annals of human genetics 72, 215-227.

24. Gresham, D., Morar, B., Underhill, P.A., Passarino, G., Lin, A.A., Wise, C., Angelicheva, D., Calafell, F., Oefner, P.J., and Shen, P. (2001). Origins and divergence of the Roma (Gypsies). The American Journal of Human Genetics 69, 1314-1331.

25. Morar, B., Gresham, D., Angelicheva, D., Tournev, I., Gooding, R., Guergueltcheva, V., Schmidt, C., Abicht, A., Lochmüller, H., and Tordai, A. (2004). Mutation history of the Roma/Gypsies. The American Journal of Human Genetics 75, 596.

26. Rai, N., Chaubey, G., Tamang, R., Pathak, A.K., Singh, V.K., Karmin, M., Singh, M., Rani, D.S., Anugula, S., and Yadav, B.K. (2012). The Phylogeography of Y-Chromosome Haplogroup H1a1a-M82 Reveals the Likely Indian Origin of the European Romani Populations. PloS one 7, e48477.

27. Mendizabal, I., Lao, O., Marigorta, U.M., Wollstein, A., Gusmao, L., Ferak, V., Ioana, M., Jordanova, A., Kaneva, R., and Kouvatsi, A. (2012). Reconstructing the Population History of European Romani from Genome-wide Data. Current Biology.

28. Altshuler, D., Brooks, L., Chakravarti, A., Collins, F., Daly, M., and Donnelly, P. (2005). A haplotype map of the human genome. Nature 437, 1299-1320.

29. Herráez, D.L., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M.R., Gross, A., Scholz, M., and Stoneking, M. (2009). Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PloS one 4, e7888.

30. Li, J., Absher, D., Tang, H., Southwick, A., Casto, A., Ramachandran, S., Cann, H., Barsh, G., Feldman, M., and Cavalli-Sforza, L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100.

31. Nelson, M., Bryc, K., King, K., Indap, A., Boyko, A., Novembre, J., Briley, L., Maruyama, Y., Waterworth, D., and Waeber, G. (2008). The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. The American Journal of Human Genetics 83, 347-358.

32. Busing, F., Meijer, E., and Leeden, R. (1999). Delete-m Jackknife for Unequal m. Statistics and Computing 9, 3-8.

33. Kunsch, H. (1989). The jackknife and the bootstrap for general stationary observations. The Annals of Statistics, 1217-1241.

34. Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics 84, 210-223.

35. Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P., Morrow, B., Friedman, E., Oddoux, C., and Burns, E. (2010). Abraham's Children in the Genome Era: Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry. The American Journal of Human Genetics.

Chapter 5

Genetic evidence for recent population mixture in India*

This work is currently under-review. The author list is: Moorjani P[+], Thangaraj K[+], Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D[§], Singh L[§]

[+] - Authors contributed equally

[§] - Authors co-directed the project

Author Contributions:

Conceived and designed the experiments: PM, NP, DR

Collected samples: KT, GP, LS

Analyzed the data: PM, DR

Contributed data: PR, MP, KT, GP, LS

Contributed analysis tools: PM, NP, PR, MP, BB, DR

Wrote the paper: PM, NP, PR, MP, KT, LS, DR

5.1 Abstract

Most Indian groups descend from a mixture of two highly divergent populations: Ancestral North Indians (ANI) related to Central Asians, Middle Easterners, Caucasians and Europeans, and Ancestral South Indians (ASI) not closely related to groups outside the subcontinent. The date of mixture is unknown but is central for understanding Indian history. We report genome-wide data from 73 groups from the Indian subcontinent and analyze linkage disequilibrium to estimate ANI-ASI mixture dates of 1,900-4,200 years ago. In at least a subset of groups 100% of the mixture is consistent with having occurred during this period. These results show that India experienced a demographic and cultural transformation several thousand years ago, from a region in which major population mixture was common, to one in which mixture even between closely related groups became rare because of a shift to endogamy.

5.2 Introduction

Nearly every group in India today descends from a mixture of two genetically divergent ancestral populations, Ancestral North Indians (ANI) related to West Eurasians (Central Asians, Middle Easterners, Caucasians and Europeans), and Ancestral South Indians (ASI) related (distantly) to indigenous Andaman Islanders[1]. Multiple lines of evidence from linguistics[2], and genetics[1; 3] have documented that West Eurasian admixture is pervasive in India and signatures of the mixture are present in all traditional caste and tribal groups, and in speakers of different language families.

Archaeological and linguistic studies in India support the presence of two distinct groups in South Asia-- one related to West Eurasians and another group that is native to India. Archaeological studies highlight that agriculture on the subcontinent developed at different time periods in the north and south of India. The earliest record of agriculture dates back to about 9000-8000 years before present from Mehrgarh in Pakistan, which included crops such as wheat and barley that were similar to the crops found in Near East during this time period[4; 5]. In contrast, agriculture developed much later in southern India around 4,600 years and predominantly included native pulses such as mungbean and horsegram[6]. Comparative linguistic studies also provide support for this dichotomy. There are about 1600 languages spoken in India. A vast majority of these, however, belong to Indo-European and Dravidian language families[2]. These languages are very distinct and appear to have very independent histories- Indo-European languages such as Sanskrit and Hindi are part of a large language family that includes Latin, English and many European and West Asian languages. On the other hand, Dravidian languages such as Tamil and Telugu are spoken by 200 million people and have been hypothesized to be native to India. In addition, the presence of loanwords (borrowed vocabulary) that trace their origin to Munda or Dravidians languages has been identified in Sanskrit and not in any other Indo-European language spoken outside India, suggestive of mixing of cultures and peoples[7; 8]. In addition, genetic studies based on Y-chromosome[9; 10], mitochondrial DNA[11; 12] and more recently autosomal DNA[3; 13] have shown beyond doubt that Indians harbor West Eurasian ancestry and individuals cluster based on ethnicity, language geography.

All of these studies have provided strong evidence for the admixture in India; however, the dates of mixture remain unknown. There are three main demographic events that could be related to the presence of West Eurasian ancestry in India: (1) Individuals of West Eurasian

ancestry could have arrived in India as part of post ice-age migrations. The earliest civilization in India, the Indus Valley civilization, dates back to 3300 BC[14]. Trade connections between Indus and Mesopotamia are well recognized[15], providing opportunity for contact between the West Eurasians and the indigenous people of India; (2) Individuals of West Eurasian ancestry could have immigrated to India with the spread of agriculture which would imply dates of either 9000 or 5000 years before present[4; 6]; (3) Finally, the appearance of Indo-Europeans languages and Vedic religion in India by 1500 BC[16; 17] could have provided an opportunity for the mixture of individuals of ANI ancestry and the then inhabitants of India (ASI).

Inferring the date can shed light on these demographic movements and improve our understanding of the formative period of Indian history.

5.3 Materials and Methods

Datasets

To learn about population history in India at higher resolution than was previously possible we assembled data for 571 South Asian individuals from 73 well-defined ethno-linguistic groups (including two Pakistani groups). For samples genotyped on Affymetrix 6.0 arrays we required at least 99% completeness for all single nucleotide polymorphisms (SNPs) and samples; this resulted in 383 individuals from 52 groups (27 groups newly genotyped for this study)[1] typed at 494,863 SNPs. For samples genotyped on Illumina 650K arrays we required at least 95% completeness yielding 188 individuals from 21 groups[3; 18] typed at 543,980 SNPs.

We initially filtered out 49 samples based on three criteria: (a) We removed one sample from each pair of duplicates (individuals that match at least 90% of genotypes); (b) we removed related individuals (for mother-father-child trios we exclude the child, and for other relative pairs

we remove one of the two individuals); (c) we removed all samples previously excluded by Metspalu et al (2011); and (d) we removed six Pakistani groups (Hazara, Kalash, Burusho, Makrani, Balochi and Brahui) as it has been previously shown that these groups have a more complex history than simple mixture of two ancestral groups[1] (Table D.1).

We next excluded an additional 194 samples based on patterns of relatedness observed in Principal Component Analysis (PCA): (1) We removed samples and groups that have evidence of recent ancestry from groups other than ANI and ASI based on previous documentation[1; 19] and PCA of Figure 5.1 which removed all Austro-Asiatic and Tibeto-Burman speakers; (2) We removed groups that were not homogenous in PCA; and (3) We removed individuals who do not cluster with the majority of samples from their own group (Table D.1).

We merged the data with Human Genome Diversity Panel (HGDP) data from 51 groups (257 individuals genotyped on Affymetrix 500K SNP array[20], and 940 on Illumina 650K array[18]); International Haplotype Map Phase 3 (HapMap3) data from 11 groups (1,158 individuals genotyped on Affymetrix 6.0 array and Illumina 1M array[21]); Behar et al. (2010) data from 41 groups (466 individuals genotyped on Illumina 610K array[22]); and Yunusbayev et al. (2011) data from 13 groups (214 individuals genotyped on Illumina 610K array[23]).

We created an "Affymetrix" dataset of 210,482 SNPs by merging data on 211 Indians (30 groups) with data from non-Indians typed on Affymetrix arrays (HapMap3 and Affymetrix HGDP). We created an "Illumina" dataset of 500,703 SNPs by merging data on 117 South Asians (15 groups) with data from non-Indians typed on Illumina arrays (HapMap3, Illumina HGDP, Behar et al (2010) and Yunusbayev et al. (2011)). We intersected these data to create an "Illumina-Affymetrix" dataset of 328 South Asians (45 groups) genotyped at 86,213 SNPs.

F_4 Ratio Estimation

We use *F_4 Ratio Estimation* as implemented in ADMIXTOOLS[24] to estimate the proportion of ANI ancestry in Indian groups. Specifically, we use the statistic:

$$ANI\% = \frac{f_4(YRI, Basque; India, Onge)}{f_4(YRI, Basque; Georgians, Onge)} \quad [1]$$

This assumes the model of Figure D.1 with Pop1 = Georgians, Pop2 = Basque.

This is different from the model used previously in Reich et al (2009) which uses Papuans, CEU and Adygei as outgroups[1]. Since the publication of Reich et al. (2009), we have observed that some of the groups used in the statistic have a more complex history than is captured by the model[25]. Hence for this study, we replace Papuans with Yoruba (YRI) and use new West Eurasian outgroups that provide a better fit to the proposed model of historical relationships.

For *F_4 Ratio Estimation* to provide unbiased results, it requires access to four outgroup populations that branch off at four distinct positions on the ancestral lineage relating ANI and ASI[24]. We choose to work with Yoruba (YRI), Andamanese (Onge)[26] and two West Eurasian populations (Pop1 and Pop2) that are at successively increasing phylogenetic distance from the ANI (that is, the tree for West Eurasian populations is (Pop2, (Pop1, ANI))) (Figure D.1). We first fixed Pop1. For each Indian group (*X*), we compute *D(Onge, X; YRI, Y)* where *Y* = any West Eurasian population from a panel of 42 groups including Europeans, Central Asians, Middle Easterners, and Caucasian populations. For all 45 Indian groups on the cline, we find that Georgians are consistent with being a clade with ANI (Table D.2, Table D.3). Thus we use Georgians as Pop 1 (Figure D.1). We next fixed Pop2. We examine all possible West Eurasian

populations to find groups that provide a good fit to the proposed phylogeny (YRI, (Pop2, (Georgians, ANI)), (ASI, Onge)) using our *admixture graph* phylogeny testing software[24]. Within the limits of our resolution, we find 5 groups (Pop2 = Italian, Tuscan, Basque, Kurd, Abhkasian) that are consistent with this model in the sense that none of the *f*-statistics relating the groups are greater than three standard errors from expectation. To obtain a sense of the uncertainty in the ANI ancestry proportions ranging over these five candidates for Pop2, we ran *F_4 Ratio Estimation* with two choices of Pop2 representing different geographic extremes (Pop2 = Abhkasian, and Pop2 = Basque). We obtain similar ANI ancestry estimates (Table D.4). Our estimates are also consistent (within 2 standard errors) to those in Reich et al. 2009 (Table D.4).

Estimating admixture dates using *rolloff*

For each pair of SNPs (*x,y*) separated by a distance *d* Morgans, we compute the covariance between (*x,y*), which we used to measure the linkage disequilibrium (LD) due to population mixture. Specifically, we use the *rolloff*[24; 27; 28] statistic:

$$R(d) = \frac{\sum_{|x-y|\approx d} z(x,y)w(x,y)}{\sum_{|x-y|\approx d} w(x,y)^2} \qquad [2]$$

where $z(x,y)$ is the covariance between SNPs *x* and *y*, and $w(x,y)$ is a weight function. The weight can be either: (a) the allele frequency difference between two groups we use as surrogates for the ancestors (Europeans, Onge); (b) the allele frequency difference between a tested Indian group and one reference (Europeans); or (c) the PCA-based SNP loadings for SNPs (*x, y*) computed by performing PCA with Europeans and various Indian cline groups. We plot the weighted covariance with distance and obtain a date by fitting an exponential with an affine

term: $y = Ae^{-nd} + c$, where d is the distance in Morgans and we interpret n as the number of generations since admixture. We compute standard errors with a weighted Block Jackknife[29], with one chromosome dropped per run.

Indo-European and Dravidian admixture dates and their difference

For our date estimates reported in Figure 5.2, we applied *rolloff* to the merged Illumina-Affymetrix dataset of 86,213 SNPs, using weights from PCA-based SNP loadings computed using Basque and all samples on the Indian cline speaking the language group other than the one being analyzed. To compute the significance of the difference in the date estimates, we leave out each of the 22 chromosomes in turn, and use a weighted Block Jackknife procedure to convert the variability in the difference into a standard error. As a robustness check, we repeated this analysis using the Affymetrix dataset of 210,482 SNPs for the 4 Indo-European groups and 5 Dravidian groups that we found were consistent with a simple ANI-ASI mixture (using Basque and all the other Indian cline groups for computing SNP loadings). We confirm a significantly younger date in Indo-European than in Dravidian speakers, with the difference of 44 ± 18 generations being significant at Z = 2.4.

Identifying groups consistent with simple ANI-ASI admixture

For each of 37 Indian groups including Onge (less than the total number of groups we had after curation because we applied a minimum sample size requirement of 5), we tested if they are consistent with deriving all their ancestry from the same ANI-ASI ancestral populations by studying the matrix of all possible statistics of the form $f_4(India_{base}, India_{other}; NonIndian_{base}, NonIndian_{other})$ comparing to a panel of 38 non-Indian populations. Many f_4 statistics can be

written as linear combinations of each other, and thus we need to pick a *basis* for the space of f_4 statistics. In practice, we pick one Indian group as "$India_{base}$" and any other Indian group (from the remaining 36 groups) as $India_{other}$. We pick an African group (YRI) as "$NonIndian_{base}$" and the $NonIndian_{other}$ groups include Dai, Papuans, Karitiana, and diverse West Eurasian groups including Europeans, Middle Easterners and Caucasians (The choice of base has no mathematical impact on the test). To identify sets of Indian groups consistent with having the same relationship to the panel of Eurasians, we use a Hotelling T-test to evaluate if the matrix of all f_4 statistics has exactly one linearly independent component (*rank 1*)[30]. For sets of Indian groups that are consistent with being rank 1, we also run the *admixture graph* software to evaluate if the relationships in Figure D.1 (where Pop 1 = Georgians, Pop 2 = Basque) are consistent with the data. We began by applying this procedure to all possible sets of three Indian groups. For the sets that passed, we added each possible fourth Indian group in turn and tested the consistency with a simple ANI-ASI mixture. We applied this process iteratively until no additional Indian groups could be added to the rank 1 set (see Note D.3 for details).

***Admixture graph* analysis**

We applied the *admixture graph*[24] software to study if the model of simple ANI-ASI admixture in rank 1 Indo-European and Dravidian groups provides a fit to the data. *Admixture graph* studies the correlations in allele frequency differentiation statistics (f_2, f_3 and f_4)[24] among groups, comparing the observed values to those specified by the model (with a standard error from a block jackknife) to test hypotheses about population relationships. To test if the model provides a fit to the data, the software examines individual *f*-statistics and considers statistics more than three standard errors from expectation to be indicative of a poor fit. We also used this

method to estimate the internal drift lengths [Figure D.2, genetic drift separating (X", ANI) and drift separating (X", ASI)] that are required by *ALDER* for estimating admixture proportions.

Estimating the date and proportion of recent admixture using *ALDER*

We run *ALDER*[31] with one reference population (we quote results from Basque, but obtain similar results with 8 other reference West Eurasian groups). The *ALDER* statistic for measuring admixture LD is similar to the *rolloff* statistic:

$$a(d) = \frac{\sum_{S(d)} z(x,y)w(x,y)}{|S(d)|} \qquad [3]$$

As before, $z(x,y)$ is the covariance between SNPs *x* and *y*. Here $w(x,y)$ is the product of the allele frequency differences at *x* and *y* between the two reference groups (in this case, Basque and the admixed group itself), and $S(d) = \{(x, y): |x - y| < d - \varepsilon/2\}$ (where ε is a discretization parameter).

We plot the weighted covariance against genetic distance and perform a least-squares fit using $y = Ae^{-nd} + c$, where *n* is the number of generations since admixture and *d* the distance in Morgans. Under a single-wave mixture model, the amplitude of admixture LD decay, defined as $a_o = A + c/2$, is analytically predicted by the ANI ancestry proportion (α) using the relationship:

$$a_o = 2\alpha(1-\alpha)(\alpha f_2(ANI,X") - (1-\alpha) f_2(ASI,X"))^2 \qquad [4]$$

Here, *X''* is the common ancestor of Basque (*X*) and the ANI-ASI lineage (Figure D.2). We estimate $f_2(ANI,X")$ and $f_2(ASI,X")$ using our *admixture graph* software with one West

Eurasian outgroup (as we do not have access to Georgians in the 210,482 SNP Affymetrix dataset). Having only a single West Eurasian outgroup in the admixture graph makes the model poorly constrained, but we can compensate by fixing the value of the admixture proportion to be equal to the ANI ancestry inferred from *F_4 Ratio Estimation.* We compare the expected amplitude a_0 (from the formula above) and the observed amplitude $\hat{a}_0$ (from the weighted LD curve) to test if the model of a single wave of mixture between ANI and ASI provides a good fit to the data (see Note D.4 for details). The entire procedure is repeated dropping out each chromosome in turn to generate block jackknife standard errors on the quantities of interest.

95% confidence interval on the ANI ancestry proportion prior to mixture

Consider the model that an Indian group derives its ancestry from two waves of admixture involving ANI-related populations (that have the same allele frequencies) and an older wave that is old enough that its contribution to the measured LD is negligible. Thus the group would have ancestry from three sources: old ANI ancestry (α_{old}), recent ANI ancestry (α_{new}) and ASI ancestry ($1-\alpha_{total}$). The expected one-reference *ALDER* amplitude is then:

$$a_o = \frac{2\alpha_{new}(1-\alpha_{total})}{\alpha_{old}+(1-\alpha_{total})}(\alpha_{total} f_2(ANI,X") - (1-\alpha_{total}) f_2(ASI,X"))^2 \qquad [5]$$

Again, we can estimate the internal drift lengths using *admixture graph* and estimate α_{total} using *F_4 Ratio Estimation*. Plugging in $\hat{a}_0$ (from the weighted LD curve) and solving the above equation for each jackknife run, we can estimate the range of $\alpha_{old.}$ We compute a one-sided 95% confidence interval of 0% to mean + 1.65 times the standard error.

5.4 Results

Data from 73 South Asian groups

We assembled the most comprehensive study of Indian genetic variation to date: genome-wide single nucleotide polymorphism (SNP) data on 571 individuals from 73 well-defined ethno-linguistic groups, including 383 individuals from 52 groups genotyped on Affymetrix SNP arrays 6.0 (27 groups newly genotyped for this study)[1] (Figure 5.1) and 188 individuals from 21 groups (including two from Pakistan) genotyped on Illumina 650K SNP arrays[3; 18]. We curated these data using Principal Component Analysis (PCA), removing individuals who did not cluster with others from the same group (Table D.1) (Methods).

Mixture proportions

Almost all groups speaking Indo-European or Dravidian languages lie along a gradient of varying relatedness to West Eurasians, as can be seen in PCA (Figure 5.1), which we have previously shown reflects variable proportions of ANI-ASI ancestry[1]. Groups speaking Austroasiatic and Tibeto-Burman languages fall away from this "Indian cline" (Figure 5.1), consistent with ancestry from distinct populations; the history of these groups is important but is not our focus here. We restricted our analysis to 45 groups that fall on the Indian cline, all of which speak Indo-European or Dravidian languages. Using *F_4 Ratio Estimation*[24] which analyzes allele frequency correlation patterns to infer mixture proportions, we estimate that the ANI ancestry ranges from as low as 17% (Paniya) to as high as 71% (Pathan) (Table D.4). Traditionally lower caste, Dravidian-speaking, and tribal groups tend to have lower proportions of ANI ancestry than traditionally upper caste and Indo-European speaking groups ($P<0.001$)[1]. Our estimates of ANI ancestry are lower than we previously reported (although within two

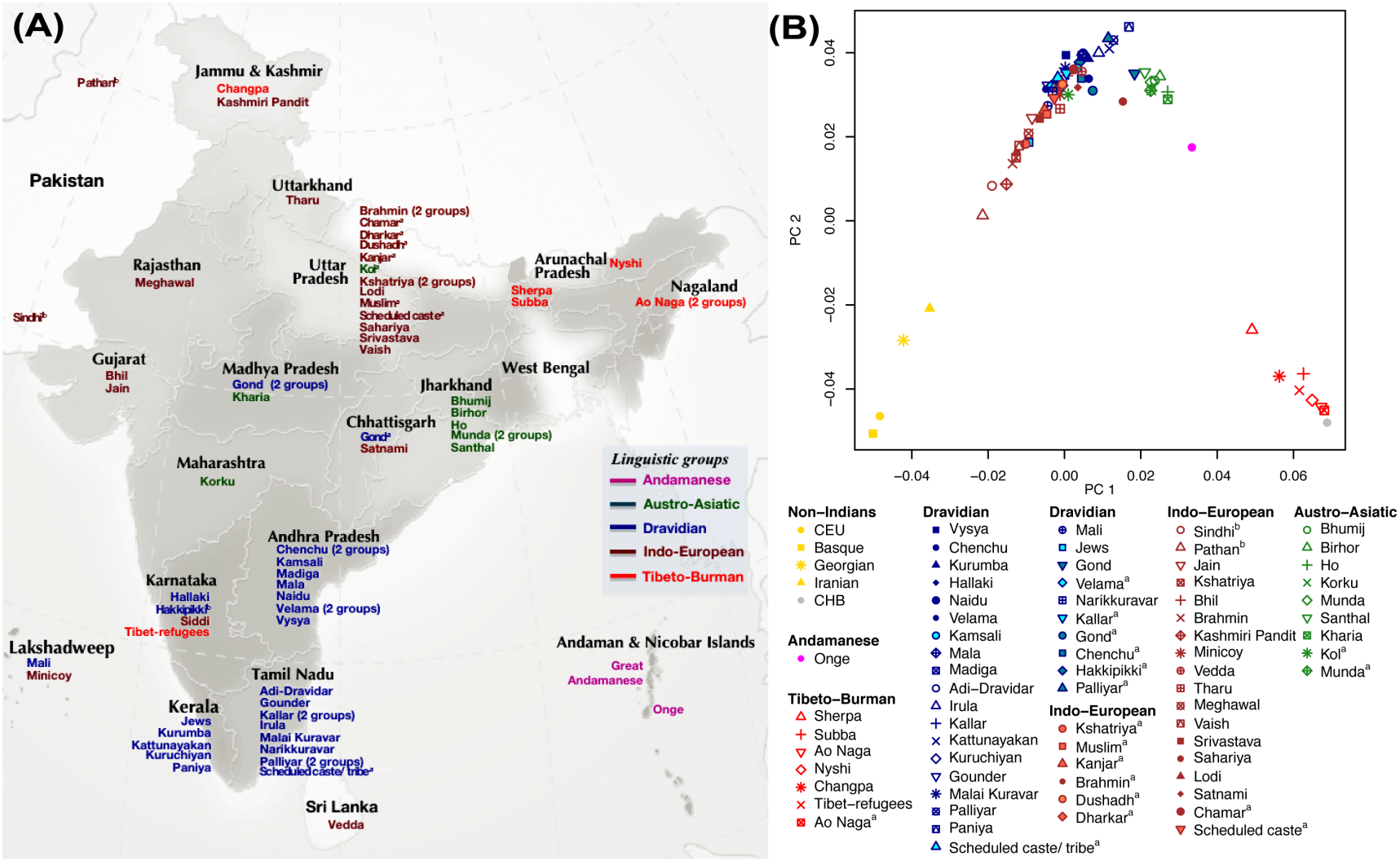


Figure 5.1. Principal Component Analysis (PCA).

Figure 5.1. (Continued) (A) Map showing the sampling locations for all Indian groups in our study (except the group labeled as "central_mix1_nihali" in [3]). **(B)** PCA of 70 of the 73 groups in this study along with non-Indians [European Americans (CEU), Georgian, Iranian, Basque, and Han Chinese (CHB)] highlights the "Indian cline," a gradient of ancestry in which northern groups (Sindhi, Pathan, and Kashmiri Pandit) show the closest relatedness to West Eurasians. To aid visualization, we represent each group by the average PCA coordinates of all the individuals in it. [a] indicates groups from Metspalu et al (2011) and [b] indicates the groups from HGDP. Great Andamanese and Siddi are not included because of their evidence of relatively recent admixture with non-South Asian groups. Also *central_mix1_nihali* from ref. [3] has not been included because it contains data from multiple ethno-linguistic groups combined under one label.

standard errors)[1], due to the fact that we previously used Papuans, Adygei, and Northwest Europeans (CEU) as outgroups for ancestry estimation, whereas here we use Yoruba, Basque and Georgians (Figure D.1, Table D.4). The reason for replacing the Papuans with YRI is that Papuans harbor gene flow from archaic humans (Denisovans)[25], which could bias ancestry estimates. We also use different West Eurasians because the Adygei derive a small proportion of their ancestry from an East Asian related source which could again bias estimates, and because a model in which Georgians are the most closely related West Eurasian group to the ANI provides a good fit to the data for many models we tested, whereas models with Europeans in their place are not as good fits. While we believe that the Onge is only distinctly related to ASI, we do not replace Onge here as they are the best surrogate population we have for ASI as most of the Indian groups on the mainland are admixed and would not be appropriate for this analysis. In addition, if the Onge harbor ancestry from groups other than ASI (although our formal tests provide evidence to the contrary) then this puts a lower bound on the estimates of admixture proportions reported here.

Admixture dates

To date ANI-ASI mixture, we capitalized on the fact that admixture between two differentiated populations generates allelic association [linkage disequilibrium (LD)] between pair of SNPs[32]. The LD decays at a constant rate as recombination breaks down the contiguous chromosomal blocks inherited from the ancestral mixing populations. The expected value of the admixture LD is related to the genetic distance between SNPs (the probability of recombination per generation between them) and the time that has elapsed since admixture[32]. We previously reported simulations showing that dating population mixture based on the scale of admixture LD is robust to the use of imperfect surrogates for the ancestral populations, fine-scale errors in the genetic map, and a history of founder event in the admixed population, and is able to provide unbiased date estimates for events up to 500 generations ago[24; 27; 28; 31] (we confirmed this using new simulations with demographic parameters relevant to India; Note D.1).

We estimated admixture dates for all the groups on the Indian cline with more than 5 samples (a minimum sample size is important for measuring LD with precision). We observe a decay of LD with genetic distance for the great majority of groups (Figure 5.2, Figure D.3). By fitting an exponential function using least squares (using *rolloff*[24; 27; 28]), our point estimates for the dates range from 64-144 generations ago, or 1,856-4,176 years assuming 29 years per generation[33,12].

We highlight two implications of these dates. First, nearly all groups experienced major mixture in the last few thousand years, including tribal groups like the Bhil, Tharu, and Paniya who might be expected to be more isolated. Second, the date estimates are typically more recent in groups speaking Indo-European languages (average of 72 generations) compared with groups speaking Dravidian languages (108 generations). A jackknife estimate of the difference is highly

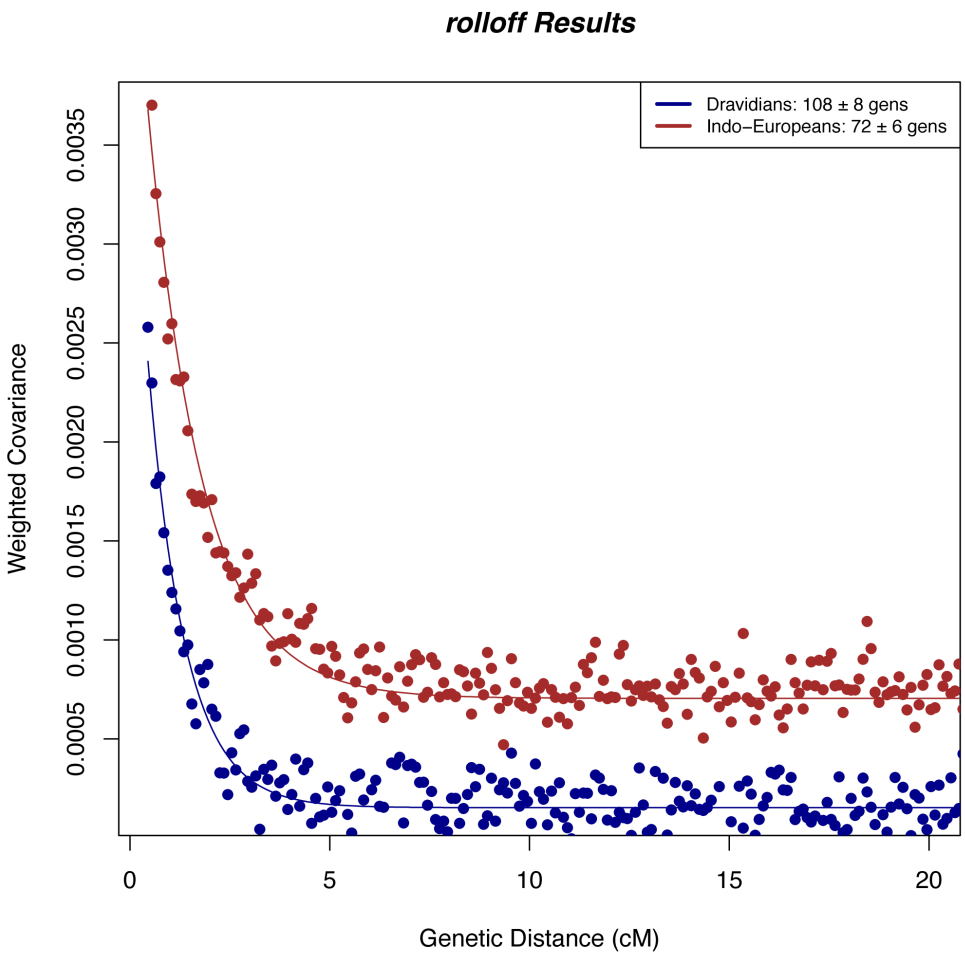


Figure 5.2. Dates of mixture. We pool samples based on linguistic affiliation (Indo-European (*n*=175) and Dravidian (*n*=144)) and run *rolloff* (using the merged Illumina-Affymetrix dataset of 86,213 SNPs) to measure the LD due to mixture between ANI and ASI. To obtain weights proportional to the allele frequency differences between ANI and ASI at each SNP (needed to run *rolloff*), we use SNP loadings obtained from a PCA of Basque and a pool of groups from the linguistic cluster whose admixture is not being dated (e.g. Indo-European when we are dating Dravidian admixture). We infer a date by fitting an exponential with a constant term, $y = Ae^{-nd} + c$, where d is the genetic distance in Morgans and where we interpret n as the number of generations since mixture. The non-zero constant (c in the fitted equation) allows for variability in the mixture proportion among the groups we pooled.

Table 5.1: Characterization of population admixture along the Indian cline

Pop	Dataset	*N*	Language-family	Traditional Caste or Social group	State/ territory	Latitude/ Longitude	ANI%	Date of admixture (gens)	Date of admixture (years)
Madiga	Reich 2009 & this study	13 (9)	Dravidian	Lower caste	Andhra Pradesh	17°58'N/79°35'N	32.0 ± 1.7	120 ± 21	3,480
Mala	Reich 2009 & this study	13 (10)	Dravidian	Lower caste	Andhra Pradesh	17°22'N/78°29'N	34.3 ± 1.7	96 ± 16	2,784
Kallar[a]	Metspalu 2011	8	Dravidian	Tribal	Tamil Nadu	10°99'N /78°22'E	37.7 ± 1.8	113 ± 15	3,277
Vysya	Reich 2009 & this study	14 (10)	Dravidian	Middle caste	Andhra Pradesh	14°41'N/77°39'N	37.9 ± 1.8	144 ± 27	4,176
Chamar[a]	Metspalu 2011	10	Indo-European	Tribal	Uttar Pradesh	25°37'N /83°04'E	38.7 ± 1.7	113 ± 13	3,277
Bhil	Reich 2009 & this study	17 (10)	Indo-European	Tribal	Gujarat	23°02'N/72°40'N	38.9 ± 1.6	78 ± 7	2,262
Scheduled caste/ tribe[a]	Metspalu 2011	6	Dravidian	Lower caste	Tamil Nadu	21°46'N /86°78'E	40.5 ± 1.9	83 ± 21	2,407
Dushadh[a]	Metspalu 2011	7	Indo-European	Lower caste	Uttar Pradesh	25°44'N /84°56'E	41.0 ± 1.8	107 ± 13	3,103
Velama[a]	Metspalu 2011	9	Dravidian	Upper caste	Andhra Pradesh	17°05'N /79°27'E	43.4 ± 1.7	85 ± 15	2,465
Dharkar[a]	Metspalu 2011	11	Indo-European	Nomadic group	Uttar Pradesh	25°44'N /83°1'E	47.8 ± 1.5	64 ± 11	1,856
Kanjar[a]	Metspalu 2011	8	Indo-European	Nomadic group	Uttar Pradesh	26°45'N /80°32'E	48.2 ± 1.7	75 ± 10	2,175
Kshatriya[a]	Metspalu 2011	7	Indo-European	Upper caste	Uttar Pradesh	27°56'N /78°65'E	54.6 ± 1.6	78 ± 9	2,262
Kshatriya	This study	15	Indo-European	Upper caste	Uttar Pradesh	25°45'N/82°41'E	60.9 ± 1.3	76 ± 10	2,204
Brahmin[a]	Metspalu 2011	8	Indo-European	Upper caste	Uttar Pradesh	26°06'N /83°18'E	61.2 ± 1.4	86 ± 7	2,494
Brahmin	This study	10	Indo-European	Upper caste	Uttar Pradesh	25°45'N/82°41'E	62.8 ± 1.4	65 ± 9	1,885
Sindhi[b]	Li 2008	10	Indo-European	Urban groups	Pakistan	24-27'N/68-70'E	64.3 ± 1.3	67 ± 8	1,943
Kashmiri Pandit	Reich 2009 & this study	15 (10)	Indo-European	Upper caste	Kashmir	34°22'N/75°50'N	65.2 ± 1.3	103 ± 17	2,987
Pathan[b]	Li 2008	15	Indo-European	Urban groups	Pakistan	32-35'N/69-72'E	70.4 ± 1.2	73 ± 9	2,117

We estimate the ANI ancestry proportion and date of admixture using F_4 *Ratio Estimation* and *rolloff* respectively, for all the groups on the Indian cline that have at least six samples (the requirement of a minimum sample size is important for measuring LD with precision). Because inferences of dates based on admixture LD are greatly improved by higher SNP density, we performed the date analysis with either the full Affymetrix (494,863 SNPs) or full Illumina (500,714 SNPs) datasets, restricting to samples typed on the relevant SNP array. For the five instances marked Reich 2009 & this study, we indicate the number of newly genotyped. [a] indicates samples from Metspalu et al (2011) and [b] indicates samples from HGDP.

statistically significant at 35 ± 8 (Z = 4.5 standard errors from zero) (Table 5.1). A possible explanation is a secondary wave of mixture in the history of many Indo-European speaking groups, which would decrease the estimated admixture date.

Table 5.2: Tests for consistency with a single pulse admixture model

Group	Language-family	Social Group	*n*	P-value
Kashmiri Pandit	Indo-European	Upper caste	15	**0.0191**
Brahmin	Indo-European	Upper caste	10	**< 0.0001**
Kshatriya	Indo-European	Middle caste	15	**0.0035**
Bhil	Indo-European	Tribal	17	**0.0010**
Vysya	Dravidian	Middle caste	14	0.0936
Madiga	Dravidian	Lower caste	13	0.0980
Mala	Dravidian	Lower caste	13	**< 0.0001**
Chamar[a]	Indo-European	Tribal	10	0.1883
Dharkar[a]	Indo-European	Nomadic group	11	**< 0.0001**
Sindhi[b]	Indo-European	Urban group	10	**0.0001**
Pathan[b]	Indo-European	Urban group	15	**< 0.0001**

Note: P<0.05 (rejection of the null model of single pulse of admixture) are highlighted in bold.

[a] indicates samples from Metspalu et al (2011) and [b] indicates samples from HGDP.

Testing for multiple layers of admixture in the history of Indian groups

A caveat for these dating analyses is that they assume that the entire admixture occurred over a short period. However, population mixture can be non-instantaneous, such that the date we obtain from our method may actually be an average of multiple dates spread out over a substantial period. One way to detect a history of non-instantaneous gene flow is to fit a sum of exponential functions to the decay of admixture LD and to show that it provides a better fit to the data than a single exponential function, as we in fact find for the Kashmiri Pandit, Kshatriya, Sindhi and Pathan (Note D.2, Table 5.2). However, even if we fail to detect a non-exponential

decay, we cannot rule out non-instantaneous gene flow, both because the decay can be noisy and because statistical detection of a mixture of exponential functions can be difficult[34]. A particularly important scenario we could not rule out using this method is that several thousand years ago, Indian groups were already admixed, and thus the LD decay we detect is the result of mixture of already admixed ancestral groups with different proportions of ANI ancestry. If the initial admixture was more than ten thousand years old, the related admixture LD would have decayed so completely as to be nearly invisible to our method. The LD we are measuring could in this case reflect only the most recent events.

To assess whether the admixture LD we are detecting could plausibly account for all the ANI-ASI mixture in an Indian group's history, we developed a novel approach: comparing the observed amplitude of the LD curves (the amount observed at short genetic distances) to what would be expected if the dated LD accounts for the entire ANI-ASI admixture. To implement this approach, we took advantage of our recently developed *ALDER* method, which computes weighted LD statistics and also provides a theoretical expectation for the amplitude under the model of a single wave of mixture, even in cases where the populations used as surrogates for the ancestral admixing populations are highly genetically drifted from the true ancestral populations[31; 35]. The *ALDER* expected amplitude formula requires estimates of the admixture proportion (which we have from *F_4 Ratio Estimation)* as well as the drift separating the true ancestral populations and our surrogates for them, which we obtain by using *admixture graph*[24] to fit a model of population relationships to the data (Methods). By comparing the observed and the expected values of the amplitude, we can evaluate whether the admixture LD we are dating can account for the entire ANI-ASI admixture in the group's history. Our simulations show that

for a single-wave admixture history, the weighted LD amplitude measured by *ALDER* is consistent with the expectation (Note D.4).

To make this analysis maximally robust, we restricted it to sets of Indian groups that are consistent with a model of mixture between the same ANI and ASI ancestral populations to within the limits of our resolution. To evaluate formally whether a model of simple ANI-ASI admixture fits the data for a proposed set of Indian groups, we compared the allele frequency differences among the Indian groups to the allele frequency differences among a set of 38 non-Indian groups including many West Eurasians, searching for differences that would be expected if the Indian groups did not derive all their ancestry from the same two ancestral populations. Specifically, we computed f_4 statistics measuring the correlation in allele frequencies between each possible pair of Indian groups in the set and diverse pairs of non-Indian groups. If the ANI ancestry in all the Indian groups in the tested set derives from the same ancestral population(s), the f_4 statistics measuring the correlations are expected to be proportional, and thus the matrix of all f_4 statistics is expected to have one linearly independent component (*rank 1*) (Note D.3). We can test this null hypothesis using a Hotelling T-test[30]. Our simulations show that this test has the power to detect a history of multiple ancestral ANI populations even when they are closely related; genetic drift in the admixed groups alone cannot increase the rank (Note D.3). For all the sets that pass as rank 1, we performed a further level of testing, running *admixture graph*[24] to evaluate if the relationships in Figure D.1 (with Georgians forming a clade with ANI and Basque as a second West Eurasian outgroup) are supported by the data in the sense that no *f*-statistic measuring allele frequency correlation is more than three standard errors from model expectation (Note D.3).

In some groups the ANI-ASI admixture is multi-layered

Applying this procedure to all possible sets of three Indian groups, and adding in additional Indian groups until we could add no more (without increasing the rank), we identified previously undetected complexity in Indian history, with many sets of Indian groups not consistent with a simple ANI-ASI admixture. This analysis produces two notable findings. First, while aboriginal Andaman Islanders (Onge) are consistent with being a sister group of ASI for many sets of Indian groups[1], the Onge cannot be added in to the model for other sets of Indian groups. Such a pattern would be expected if there was an ancient second wave of migration into the Andaman Islands from a population more closely related to the ASI ancestors of some present-day Indian groups than others; this would also be consistent with the finding that the closest matches to Andamanese mitochondrial DNA haplotypes in Eurasia are rare haplotypes found in India[36]. Second, we find that the Indian groups consistent with simple ANI-ASI mixture are most often from tribal and traditionally lower caste groups. Middle and upper caste groups tend to have evidence of more complex histories, with signals of multiple layers of ANI ancestry from slightly different ANI ancestral populations. Further evidence for multiple waves of admixture in the history of many traditionally middle and upper caste groups (as well as Indo-European speaking and northern groups) comes from the more recent admixture dates we observe in these groups (Table 5.1), and the fact that a sum of two exponential functions often produces a better fit to the decay of admixture LD than a single exponential (as noted above for some northern groups; Note D.2). This is also in agreement with the signal observed in the ADMIXTURE analysis reported in Metspalu et al (2011).

In some groups the ANI admixture is consistent with being simple and all due to events in the last few thousand years

Focusing on the largest set of Indo-European speaking groups (4 groups) and the largest set of Dravidian speaking groups (5 groups) consistent with mixture of the same ANI and ASI ancestral populations, we find that the expected and observed admixture LD amplitudes are equivalent to within the limits of our resolution. We restricted this analysis to Indian groups genotyped on Affymetrix arrays because this allowed us to analyze about 2.5 times more SNPs (n=210,482), which improves the accuracy of inferences based on admixture LD. Restricting to samples genotyped on Affymetrix arrays raised the challenge that we could not use Georgians as part of our *admixture graph* fitting (we need a second outgroup like Georgians to obtain tight constraints on the absolute estimates of ANI-ASI admixture), but in Note D.4 we show that we can accurately infer the difference between the two amplitude values (observed - expected) even without access to Georgians by constraining the admixture proportions estimated using F_4 *Ratio Estimation*. For both the Indo-European and Dravidian sets, the observed amplitudes are statistically consistent with the expected values (Table 5.3). Thus, our data are consistent with all of the ANI ancestry in selected sets of Indian groups speaking both Indo-European and Dravidian languages being due to admixture events that we can date to within the past few thousand years. Accounting for statistical uncertainty, we estimate that the ANI ancestry that cannot be explained by a single wave of admixture in the last few thousand years has a 95% confidence interval (truncated to 0) of 0-19% for the Indo-European speakers and 0-16% for the Dravidian speakers. Thus all the ANI ancestry in some groups is consistent with deriving from admixture events that have occurred in the past few thousand years.

Table 5.3. Consistent estimates of the amplitude of admixture LD for the Indo-European and Dravidian speaking rank 1 groups

Indo-European speaking rank 1 groups:			
Reference West Eurasian (X)	**Expected amplitude x 10000**	**Observed amplitude x 10000**	**Z-score for difference**
Basque	0.7 ± 0.2	0.6 ± 0.1	-0.5
CEU	0.6 ± 0.2	0.5 ± 0.1	-0.8
French	0.9 ± 0.2	0.8 ± 0.1	-0.6
Italian	0.5 ± 0.2	0.6 ± 0.1	0.1
Orcadian	0.8 ± 0.2	0.7 ± 0.1	-0.5
Sardinian	0.7 ± 0.2	0.7 ± 0.1	0.4
Tuscan	0.7 ± 0.2	0.7 ± 0.1	-0.2
Dravidian speaking rank 1 groups:			
Basque	1.1 ± 0.2	0.8 ± 0.1	-1.7
CEU	0.9 ± 0.1	0.5 ± 0.1	-2.7
French	1.1 ± 0.2	0.6 ± 0.1	-2.4
Italian	0.8 ± 0.1	0.8 ± 0.2	-0.1
Orcadian	0.9 ± 0.1	0.3 ± 0.4	-1.6
Sardinian	0.9 ± 0.2	0.9 ± 0.2	0.0
Tuscan	1.0 ± 0.2	1.0 ± 0.1	0.2

Note: We use equation [4] to compute the expected amplitude of admixture LD. The observed amplitude is based on *ALDER* analysis using X as the reference population. We ignore inter-SNP distances less than threshold chosen by *ALDER* after comparing shared LD between the reference and the admixed group. To estimate the error, we perform a weighted block jackknife, removing one chromosome in each run.

The *admixture graph* ANI admixture proportion is the inferred ANI proportion averaged over all admixed individuals in the group. The estimates for each group are weighted by their sample size.

5.5 Discussion

Our analysis provides evidence for major mixture between populations with very different ancestries in India ~1,900-4,200 years ago, well after the establishment of agriculture. We have further shown that groups with unmixed ANI and ASI ancestry were plausibly living in

India until this time. This contrasts with the situation today in which all groups in India are admixed. The major mixture we have dated is striking in light of the endogamy that has characterized many groups in India since the time of mixture. For example, the Vysya from Andhra Pradesh have experienced negligible gene flow from neighboring groups in India for an estimated three thousand years[1]. Thus, we have shown that India experienced a demographic and cultural transformation, shifting from a region where major mixture between groups was common and affected even isolated tribes such as Palliyar and Bhil, to a region in which mixture was rare.

The archaeological and historical correlates of the time of mixture are important and interesting. The period of around 1,900-4,200 years ago was a time of major change in the subcontinent, characterized by the deurbanization of the Indus valley civilization[40], repopulation of the Gangetic plateau[41], dramatic shifts in burial practices[42], and likely appearance of Indo-European languages and Vedic religion in India[16; 17]. Hints of the cultural shift towards endogamy can be observed in the evolution of ancient Indian texts. The bulk of the Rig Veda, the oldest texts composed in India, had no mention of the class or caste system, and indeed there is linguistic and philological evidence from the older part of the Rig Veda that there was acceptance of some of the pre-Indo-Aryan population as kings (or chieftains) and poets[38]. The traditional four class (varna) system, made up of Brahmanas, Ksatriyas, Vaisyas and Sudras, was first mentioned in the appendix (book 10) and was merely described as a means of social organization[37; 38]. However, assigning caste (jati) related to an individuals traditional hereditary occupation appeared only some centuries later, such as in the law code of Manu (Manusmriti), which redefined the system by forbidding intermarriage between groups and preventing the movement of individuals across caste groups[39]. Thus, over the course of the composition of the

Vedas, there is supporting evidence for India transforming from a region where mixture between divergent groups was accepted to one in which endogamy was advocated.

It is equally important to point out what our study has not shown. It is important to recognize that our results provide no direct evidence at all for people of ANI or ASI ancestry migrating to India from outside the subcontinent ~1,900-4,200 years ago. Indeed, studies that have tried to identify the West Eurasian populations that are most closely related to the main ANI ancestry component in India (as many upper and middle caste groups have multiple layers of ANI ancestry) have failed to find any population that has evidence for shared ancestry within the past 12,500 years[3], although it is possible that by surveying additional populations or deploying new methods such relatedness could be uncovered. An alternative scenario that is consistent with our results is that the ANI and ASI coexisted in India for thousands of years prior to mixing. While this hypothesis might seem unlikely at first, ancient DNA studies from Europe suggest that such scenarios are plausible and in fact directly supported by the data. In Germany and Sweden, there is converging evidence from archeology and genetics that Neolithic farming populations arrived in northern Europe around 5,000-7,500 years ago but did not admix with the local hunter gatherer groups until about 4,500 years before present. The present day genetic structure in Europe thus represents a mixture of Neolithic and Mesolithic ancestries[43-46].

In conclusion, we have documented a major cultural and demographic event that occurred in India in the period 1,900-4,200 years ago, which profoundly affected the genetic ancestry of essentially every Indo-European and Dravidian speaking population in India. An open question is the historical origin of ANI and ASI ancestry that is present not just in traditionally upper caste groups, but also in traditionally lower caste and tribal groups, all of whom are united in their history of mixture within the past few thousand years. A priority of

future studies should be to investigate ancient human remains from India, as this may make it possible to disentangle the two hypotheses consistent with our data and provide novel insights about the history of South Asia.

Acknowledgements

We thank the volunteers who donated DNA samples. We acknowledge the help of Rakesh Tamang, Justin Carlus and A. Govardhana Reddy in collection and handling of the population samples. We thank Richard Meadow and Michael Witzel for discussions about the historical contexts for these findings. P.M., N.P. and D.R. were supported by NIH grant GM100233 and NSF HOMINID grant 1032255. M.L. and P.L. were supported by NSF Graduate Research Fellowships. K.T. was supported by a UKIERI Major Award (RG-4772) and the Network Project (GENESIS: BSC0121) fund from the Council of Scientific and Industrial Research, Government of India. L.S. was supported by a Bhatnagar Fellowship grant from the Council of Scientific and Industrial Research of the Government of India, and by a J.C. Bose Fellowship from Department of Science and Technology, Government of India.

5.6 References

1. Reich, D., Thangaraj, K., Patterson, N., Price, A., and Singh, L. (2009). Reconstructing Indian population history. Nature 461, 489-494.

2. Southworth, F.C. (2005). Linguistic archaeology of South Asia.(Routledge).

3. Metspalu, M., Romero, I.G., Yunusbayev, B., Chaubey, G., Mallick, C.B., Hudjashov, G., Nelis, M., Mägi, R., Metspalu, E., and Remm, M. (2011). Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. The American Journal of Human Genetics 89, 731-744.

4. Renfrew, C. (1990). Archaeology and language: the puzzle of Indo-European origins.(Cambridge Univ Pr).

5. Costantini, L. 1984. The Beginning of Agriculture in the Kachi Plain: the Evidence of Mehrgarh, in B. Allchin (ed.) *South Asian Archaeology 1981*, pp. 29-33. Cambridge: Cambridge University Press.

6. Fuller, D.Q. (2011) Finding plant domestication in the Indian subcontinent. *Current Anthropology* 52(S4), S347-362.

7. Witzel, M. (1999). Substrate Languages in Old Indo-Aryan (Rgvedic, Middle and Late Vedic). Electronic Journal of Vedic Studies 5, 1-67.

8. Adams, D.Q. (1997). Encyclopedia of Indo-European Culture.(Routledge).

9. Kivisild, T., Rootsi, S., Metspalu, M., Metspalu, E., Parik, J., Kaldma, K., Usanga, E., Mastana, S., Papiha, S., and Villems, R. (2003). The genetics of language and farming spread in India. Examining the farming/language dispersal hypothesis, 215-222.

10. Thangaraj, K., Chaubey, G., Singh, V.K., Vanniarajan, A., Thanseem, I., Reddy, A.G., and Singh, L. (2006). In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup'M'in India. BMC genomics 7, 151.

11. Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., Dey, B., Roy, M., Roy, B., and Bhattacharyya, N. (2003). Ethnic India: a genomic view, with special reference to peopling and structure. Genome Research 13, 2277.

12. Kivisild, T., Bamshad, M.J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W.S., and Dixon, M.E. (1999). Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Current Biology 9, 1331-1334.

13. Brahmachari, S., Majumder, P., Mukerji, M., Habib, S., Dash, D., Ray, K., and Bahl, S. (2008). Genetic landscape of the people of India: a canvas for disease gene exploration. J Genet 87, 3-20.

14. Kenoyer, J.M. (1998). Ancient cities of the Indus valley civilization. (Oxford University Press Karachi).

15. Marshall, J. (1997). Mohenjo-Daro and the Indus civilization: Being an official account of archaeological excavations at Mohenjo-Daro carried out by the government of India between the years 1922 and 1927.(Asian Educational Services).

16. Trautmann, T.R. (2005). The aryan debate.(Oxford University Press).

17. Bryant, E.F., and Patton, L.L. (2005). The Indo-Aryan controversy: evidence and inference in Indian history.(RoutledgeCurzon).

18. Li, J., Absher, D., Tang, H., Southwick, A., Casto, A., Ramachandran, S., Cann, H., Barsh, G., Feldman, M., and Cavalli-Sforza, L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100.

19. Shah, A.M., Tamang, R., Moorjani, P., Rani, D.S., Govindaraj, P., Kulkarni, G., Bhattacharya, T., Mustak, M.S., Bhaskar, L., and Reddy, A.G. (2011). Indian siddis: African descendants with Indian admixture. The American Journal of Human Genetics 89, 154-161.

20. Herráez, D.L., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M.R., Gross, A., Scholz, M., and Stoneking, M. (2009). Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PloS one 4, e7888.

21. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S., Yu, F., Bonnen, P.E., De Bakker, P., Deloukas, P., and Gabriel, S.B. (2010). Integrating common and rare genetic variation in diverse human populations. Nature 467, 52.

22. Behar, D., Metspalu, E., Kivisild, T., Achilli, A., Hadid, Y., Tzur, S., Pereira, L., Amorim, A., Quintana-Murci, L., and Majamaa, K. (2006). The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. The American Journal of Human Genetics 78, 487-497.

23. Yunusbayev, B., Metspalu, M., Järve, M., Kutuev, I., Rootsi, S., Metspalu, E., Behar, D.M., Varendi, K., Sahakyan, H., and Khusainova, R. (2011). The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Molecular biology and evolution 29, 359-365.

24. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. Genetics 192, 1065-1093.

25. Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M.R., Pugach, I., Ko, A.M.S., Ko, Y.C., Jinam, T.A., and Phipps, M.E. (2011). Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. The American Journal of Human Genetics 89, 516-528.

26. Abbi, A. (2009). Is Great Andamanese genealogically and typologically distinct from Onge and Jarawa? Language Sciences 31, 791-812.

27. 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 Genetics 7, e1001373.

28. Moorjani, P., Patterson, N., Loh, P.-R., Lipson, M., Kisfali, P., Melegh, B.I., Bonin, M., Kádaši, Ľ., Rieß, O., Berger, B., et al. (2013). Reconstructing Roma history from genome-wide data. PloS one 8, e58633.

29. Busing, F., Meijer, E., and Leeden, R. (1999). Delete-m Jackknife for Unequal m. Statistics and Computing 9, 3-8.

30. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., and Mesa, N. (2012). Reconstructing Native American population history. Nature 488, 370-374.

31. Loh, PR., 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.

32. Chakraborty, R., and Weiss, K. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proceedings of the National Academy of Sciences 85, 9119.

33. Fenner, J. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology 128, 415.

34. Osborne, M., and Smyth, G. (1986). An algorithm for exponential fitting revisited. Journal of Applied Probability, 419-430.

35. Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S.W., Nakagawa, H., Naumann, C., et al. (2012). The genetic prehistory of southern Africa. Nature Communications advance online publication: doi: 101038/ncomms2140

36. Barik, S., Sahani, R., Prasad, B., Endicott, P., Metspalu, M., Sarkar, B., Bhattacharya, S., Annapoorna, P., Sreenath, J., and Sun, D. (2008). Detailed mtDNA genotypes permit a

reassessment of the settlement and population structure of the Andaman Islands. American Journal of Physical Anthropology 136, 19-27.

37. Talageri, S. The Rigveda: A Historical Analysis. 2000. In. (ISBN 81-7742-010-0.

38. Witzel, M. (1995). Early Indian history: Linguistic and textual parameters. The Indo-Aryans of ancient South Asia: Language, material culture and ethnicity, 85-125.

39. Naegele, C.J. (2008). History and Influence of Law Code of Manu.

40. Meadow, R.H., ed. (1991). Harappa Excavations 1986-1990: A Multidisciplinary Approach to Third Millenium Urbanism.(Prehistory Pr).

41. Lawler, A. (2008). Unmasking the Indus. Indus collapse: the end or the beginning of an Asian culture? Science (New York, NY) 320, 1281.

42. Sarkar, S.S. (1964). Ancient Races of Baluchistan, Panjab, and Sind.(Bookland).

43. Haak, W., Forster, P., Bramanti, B., Matsumura, S., Brandt, G., Tänzer, M., Villems, R., Renfrew, C., Gronenborn, D., and Alt, K.W. (2005). Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. Science 310, 1016-1018.

44. Bramanti, B., Thomas, M., Haak, W., Unterlaender, M., Jores, P., Tambets, K., Antanaitis-Jacobs, I., Haidle, M., Jankauskas, R., and Kind, C.-J. (2009). Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science 326, 137-140.

45. Malmer, M.P. (2002). The Neolithic of south Sweden: TRB, GRK, and STR. (Royal Academy of Letters).

46. Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M.T.P., Götherström, A., and Jakobsson, M. (2012). Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336, 466-469.

47. Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337.

Chapter 6

Conclusions and Perspectives

6.1 Contribution of this thesis

In this dissertation, I have presented a range of methods that study the distributions of allele frequency, linkage disequilibrium, and identity-by-descent sharing among individuals, for learning about history from genetic data. I have applied these methods to reconstruct several episodes of South Asian and West Eurasian history. The insights from these analyses are complementary to the inferences from other disciplines such as linguistics, archaeology and history.

In the first study (Chapter 2), I introduced a novel method for inferring the time of population admixture, and through simulations verified that this method is applicable for dating admixture events that occurred up to 300 generations (~9,000 years) before present. I analyzed genomewide single nucleotide polymorphism (SNP) data from 40 West Eurasian groups and showed that Southern European, Levantine and Jewish groups have inherited ~1-15% sub-Saharan African ancestry within the past 100 generations. The dates of admixture in Southern European and Levantine populations are consistent with events during the Roman Empire and subsequent Arab migrations. The dates in the Jewish populations are older (~72 generations ago), but have overlapping intervals in 8 diverse Jewish groups, potentially reflecting descent from a common ancestral population that already had some African ancestry prior to the Jewish Diasporas. The signal of sub-Saharan African ancestry in these West Eurasian groups in the last few thousand years confirms that there were continued contacts between Mediterranean civilizations and Africa through trade and slave expeditions, after the initial migration out of Africa.

In the second study (Chapter 3), I investigated the genomewide ancestry of Siddi groups and showed that these populations trace their ancestry to African, Indian and Portuguese ancestral groups within the past 200 years. As this mixture had occurred in the recent past, I was able to verify the consistency of the genetic inferences with the results from other disciplines, namely historical records documenting the arrival of Siddis in India. The study on Indian Siddi groups illuminates the history of the African Diaspora across the Indian Ocean, and confirms that the African slave trade extended beyond Europe, Middle East and the Americas.

In the third study (Chapter 4), I investigated the ancestry of European Roma gypsies who have been suggested to have migrated to Europe from South Asia. I performed formal tests to confirm that Roma have both West Eurasian and South Asian ancestry, and estimate that they harbor 80% West Eurasian ancestry deriving from a combination of European and South Asian sources. The examination of the patterns of identify-by-descent sharing between the Roma and non-Roma groups showed that the West Eurasian and South Asian ancestry in Roma likely traces its origin to Eastern Europeans and Northwest Indians respectively. Finally, I estimated that the major West Eurasian gene flow in to Roma occurred about 850 years ago, soon after the exodus of Roma out of the Indian sub-continent. The migration out of India was accompanied by a severe founder event, signatures of which have been preserved for hundreds of years because of the endogamy that is prevalent in Roma communities.

Finally, in Chapter 5, I investigated the history of South Asian populations that have ancestry from two highly divergent ancestral groups, Ancestral North Indians (ANI) related to West Eurasians and Ancestral South Indians (ASI) not closely related to groups outside the subcontinent. By studying admixture linkage disequilibrium (LD) in 73 South Asian groups, I estimated that major gene exchange occurred ~2,000-4,000 years before present. I developed a

novel approach that combines the insights of allele frequency and LD-based statistics to infer the underlying model of population mixture that differentiates between models of single and multiple pulses of gene exchange. I showed that almost all upper and middle caste groups have a history of multiple episodes of ANI-related gene exchanges. Thus, the date of mixture in these groups should be interpreted as the weighted average of the date of the multiple events. In a subset of tribal and lower caste groups, the formal tests indicated that all the mixture occurred during this period. However, this result was striking in the light of the endogamy that has characterized most Indian groups since the mixture. This study has shown that India experienced a major cultural transformation, changing from a region where mixture was pervasive to a region where mixture even among populations located geographically close to each other was rare.

Historical and archaeological records have suggested that this was a period of major changes in the subcontinent characterized by the downfall of the Indus civilization and the likely appearance of Indo-European languages in India. However, I caution that population mixture dates should not be interpreted as migration dates without supporting evidence from other disciplines such as linguistics and archaeology. The alternate hypothesis that the ANI and ASI populations coexisted in India long before mixing is also entirely consistent with the results of this study. One of the most important questions in Indian history is to understand the nature of events that led to the spread of Indo-European languages and Vedic culture to India, and this study has provided new insights for understanding this formative period of Indian history.

In conclusion, this thesis provides valuable insights into the history of West Eurasian and South Asian populations and helps address important open questions related to the history and ancestry of these groups.

6.2 A limitation in the precision of the dates

In this study, the time of mixture (in generations) was inferred by studying the decay of admixture linkage disequilibrium in the target population. To understand the demographic movements related to the mixture and to compare the estimates across disciplines like linguistics and archaeology, the date in years is more useful. I computed the date estimate in years by multiplying the date estimate in generations and the generation interval (average number of years per generation). The generation interval has been estimated to be 29 years in diverse populations including hunter-gatherer, developing, and industrialized nations[1; 2]. It depends on a range of factors such as gender, culture, geography and time; and the point estimate is likely not capturing all the uncertainty in this parameter[1]. However, the effect of each of these factors is not well characterized. As more information related to the human reproductive behaviors becomes available, it may become possible to estimate the changes in generation interval related to each of these factors, and compute more precise estimates of the dates of mixture in years.

6.3 Future directions

In this section, I outline three main directions of investigation that I believe will be important to pursue in terms of using genetic data to provide new insights about the history of groups investigated in this study. Because a discussion about the future directions has been included at the end of each chapter, this section is relatively broad, as it is applicable to most studies of human population history.

Broadening sample coverage

A key area for future work should be to identify the exact (or the closest) population that has contributed to the admixture (often referred to as the source of the admixture) as this can provide important insights about the evolutionary history of populations. Such studies would greatly benefit from improving the spatial coverage of samples across the world, in particular Africa and East Asia, where the current sampling is sparse. Comprehensive coverage of samples from Africa will help identify the source of the sub-Saharan ancestry in South Europeans, Middle Easterners and Jewish groups.

Recent work in extracting and sequencing of ancient human remains has provided, for the first time, an opportunity to directly sample human variation from the past. Most ancient DNA studies have so far focused on Europe, as the climate preserves samples relatively well there, but it would be very exciting to sequence samples from South Asia and Africa, even if these studies are limited to sampling genetic variation from the past few thousand years. In order to trace the historical origin of the ANI-ASI populations, it would be of particular significance to sequence ancient human remains from Bronze-age sites in South Asia, such as Harappa and Mohenjo-daro (in Pakistan)[3; 4]. The ancestry of the occupants of these sites remains unclear, but uncovering this information provides an opportunity to improve our understanding of the history of South Asia; and also the origin and spread of Indo-European languages across the world.

Power of sequence data

Technological and scientific advances have made sequencing the entire genomes of large numbers of individuals feasible and cost-effective. This provides an exciting opportunity to

analyze ascertainment-bias free data that samples all the variation in the genome, including mutations that are only found in the person being sequenced (private alleles). This provides an opportunity to make inferences of demographic parameters based on the allele frequency spectrum, without modeling or correcting for complicated ascertainment schemes. In addition, examining the distribution of rare alleles can provide information about recent population structure. The identification of shared haplotype segments containing rare or private alleles can serve as a signature of ancestry in admixed populations and studying the geographic sharing can help uncover the source of the admixture. Furthermore, the distribution of the length of the haplotypes can provide important information about the timing of demographic events such as population mixtures and divergences.

Developing new statistical methods

A particularly exciting area of method development involves combining inferences from modern and ancient DNA samples. Integrating the patterns of spatial and temporal distributions can provide new insights about the evolutionary history of populations. For example, current studies investigating spatial relationships of European populations have been unable to resolve if the clinal structure observed in principal component analysis of Europeans is driven by isolation-by-distance or by historical migrations or both. Methods that can characterize the evolution of spatial population relationships over time would, however, be able to disentangle the signals of genetic drift and migrations, thus providing a way to differentiate between these two hypotheses and characterize the contribution of each.

Another important topic to investigate is "branch shortening" on the archaic lineage. Depending on the samples age, the archaic sample could lack several thousands of years of molecular evolution compared to a modern sample. Thus by measuring the divergence of archaic and modern samples from a shared common ancestor such as chimpanzee, one can estimate the age of the archaic fossil. This insight was used by Meyer et al. (2012) to estimate the date of the Denisova fossil[5]. However, this requires calibration to human-chimpanzee divergence time that is not very accurately determined. An alternate approach would be to study patterns of shared LD between ancient and modern samples, perhaps related to a shared gene flow or founder event, to characterize the relative dates of the two samples. The LD based approach does not require calibration and can thus provide unbiased estimates. Inferences of branch shortening can have a wide range of applications from estimating the age of fossils to understanding population divergence times to calculating the mutation rate in humans.

6.4 References

1. Fenner, J. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology 128, 415.

2. Helgason, A., Hrafnkelsson, B., Gulcher, J.R., Ward, R., and Stefánsson, K. (2003). A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. The American Journal of Human Genetics 72, 1370-1388.

3. Meadow, R., and Kenoyer, J. (2001). Excavations at Harappa 2000-2001: New insights on chronology and city organization. South Asian archaeology, 207-226.

4. Possehl, G.L. (2002). The Indus civilization: A contemporary perspective.(Altamira Pr).

5. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-226.

Appendix A

Supplementary Material for Chapter 2

Text A.1: PCA-based search for outliers and sub-structure

We performed Principal Component analysis (PCA) analysis with the unmixed ancestral West Eurasians (CEU) and ancestral Africans (YRI) and each West Eurasian population (*X*) to study the relationship of individuals within population *X* to sub-Saharan Africans. A plot of the first and second PCs separates the YRI from CEU and West Eurasians fall between these two populations, depending on their genetic proximity to Africans (Figure A.1). We observe that most populations are homogenous in their relationship to sub-Saharan Africans, with a few outliers. In order to obtain sample estimates that are more representative of the bulk of the populations, a total of 140 outliers were removed based on PCA (Table A.2). Here are a few notes related to the PCA.

1. We removed all samples from the POPRES Greece population from our analysis, as PCA showed that the 8 Greek individuals formed 3 separate clusters, with evidently different proportions of African relatedness. It was not clear which cluster was most characteristic of the larger Greek population, and hence we excluded the population altogether.

2. Combined PCA of CEU, YRI and all Italian populations shows that there are three significant clusters among the Italian populations: Sardinian, Northern-Italy and Southern-Italy. The "Southern-Italy" group mainly consists of individuals from the POPRES Italy population that appear to be sampled from the south of Italy based on the clustering seen in PCA. The "Northern-Italians" group contains data from POPRES Italy and Swiss-Italians and HGDP-CEPH Bergamo and Tuscany and "Sardinians" include individuals from HGDP-CEPH Sardinia and POPRES Italy population, which appear to be closely related to this group (Figure A.1C). Individuals that did not fall into these three main clusters were excluded from all further analysis.

3. The HGDP-CEPH Bedouin population was divided into 2 groups – Bedouin-g1 and Bedouin-g2. The Bedouin-g1 population is more similar to the HGDP-CEPH Palestinian population (Figure A.1D) compared to the other Bedouin group (Bedouin-g2).

4. PCA analysis of YRI, CEU and 389 IBD Ashkenazi Jews shows that a large number of Ashkenazi Jews cluster together forming a main groups of ~320 individuals. The rest of the individuals are very heterogeneous and we hypothesize that they have experienced recent admixture with non-Ashkenazi groups (Figure A.1E). We excluded 69 individuals that do not fall in the main cluster.

5. In the case of the Jewish samples from the Jewish HapMap project, we applied a slightly different algorithm to search for outliers. In addition to performing PCA for each Jewish population *X* and CEU and YRI, we performed a combined PCA of all Jewish populations. This was done to confirm that the self-reported ancestry of the Jewish individuals correlates with the ancestry based on genetic data. In some cases, we found that there was significant heterogeneity in the population (for example: Italian Jews). Hence, we removed all individuals that did not cluster with the bulk of individuals of that group. We also identified some individuals that clustered with groups other than their own, and these individuals were also excluded from further analysis (Figure A.1E).

Comparison of results before and after curation shows that the data curation does not affect the qualitative inferences (Table A.3).

Text A.2: Robustness of inferences to the choice of ancestral populations

To test whether our inferences are sensitively dependent on the ancestral populations used as reference populations, we performed all the tests substituting YRI and CEU with other related populations.

(a) We computed principal components (PCs) with EIGENSOFT [2] using the Kenyan Bantu and Adygei (a population from the Northern Caucasus) and projected various West Eurasian and African populations onto the PCs. A plot of the first and second PCs shows that West Eurasian populations form a gradient of relatedness to Kenyan Bantu, with other African populations having frequencies that are most correlated to Kenyan Bantu and the Northern European populations having frequencies that are least correlated. We observe that Southern European, Jewish and Levantine populations are close to Adygei but are slightly shifted towards Kenyan Bantu compared to Northern European populations, just as in Figure 2.1 (Figure A.2).

(b) To assess the robustness of our inferences from Table 2.1, we repeated the *4 Population Test* with alternate tree topologies where we replaced the ancestral West Eurasian and ancestral sub-Saharan African populations. Alternate topologies tested were as follows:

(i) ((YRI, Papuan), (Adygei, *X*)) (*X* is a range of West Eurasian populations tested)

(ii) ((Mandenka, Papuan), (CEU, *X*))

(iii) ((Kenyan Bantu, Papuan), (CEU, *X*))

In all three cases, we observed significant violations of the expected tree topology for Southern European, Jewish groups and Levantine populations (Table A.4). The Northern European populations showed no evidence of admixture (Table A.4).

(c) To demonstrate that our estimates of African mixture proportions are not sensitive to the sub-Saharan African group we compared to, we use *f_4 Ancestry Estimation* [3] with the alternate phylogenetic tree (San,(*Y*,(Papuan,(CEU,*X*)))), where *X* = range of West Eurasian populations that show violations of the *4 Population Test* [3] and *Y*= Mandenka or Kenyan Bantu. We also replaced the outgroup Papuan with the HapMap3 Han Chinese (CHB). Our analyses show that the tests are robust to the ancestral populations or the outgroup chosen, as the results are qualitatively similar.

Text A.3: Effect of SNP ascertainment bias on results of *3 Population Test*

To investigate if SNP ascertainment bias can affect the results of the *3 Population Test*, we performed coalescent simulation using Hudson's *ms*[4] to generate data for two ancestral populations, Population *A* (Pop *A*) and Population *B* (Pop *B*). For the simulation, we use a two-population demography similar to one used in reference[1] where the effective population size of Pop A is N_0 = 10,000 and effective population size of Pop B varies from $0.25N_0$ to $0.85N_0$ such that the frequency differentiation F_{ST}(A,B)=0.15 and the divergence time varies from 45,000 - 100,000 years. These simulated populations can be roughly considered as Africans and Europeans. Details of the demographic model are presented in Figure A.4. In order to generate SNPs affected by ascertainment bias, we select two chromosomes and examine the alleles. If the two alleles are different, we record the data for the SNP. There are three ascertainment schemes:

(1) One chromosome is selected from each of Pop *A* and Pop *B*

(2) Both the chromosomes are selected from Pop *A*

(3) Both the chromosomes are selected from Pop *B*

We construct the genomes of individuals of Pop *C* who have mixed Pop *A* and Pop *B* ancestry by using the simulator described in the Materials and Methods section. We choose a 20%/80% mixing proportion and set the time since mixture to be 10 generations. We then perform the *3 Population Test*[3] (C; A, B). We observe that regardless of the ascertainment scheme or the demographic model chosen, there is clear evidence of mixture in Pop *C* (Table A.6) and our results are not affected by ascertainment bias.

Text A.4: Robustness of the *ROLLOFF* method for estimating mixture dates

(a) *ROLLOFF* simulations for a scenario similar to African Americans

We simulated genomes of 10 individuals of mixed European and African ancestry using the simulation framework described in the *Materials and Methods*. We set the time since mixture (λ) at 6 generations and the European ancestry proportion (θ) was sampled from a beta distribution with mean 20% and standard deviation 10%. These parameters were chosen to be within the ranges of values that are typical for African Americans[5]. *ROLLOFF* analysis was performed using a non-overlapping dataset of 1,107 European American and 737 Yoruba Nigerian individuals as reference populations. The analysis was restricted to 339,171 SNPs and the fine scale recombination map of Myers et al.[6] was used for specifying the genetic distance.

A plot of the admixture LD shows an approximately exponential decay of LD with genetic distance. The half-life suggests a mixture date of 6 ± 1 generation (Figure A.7). The standard error was calculated using the *Weighted Block Jackknife* where we remove one

chromosome in each run and measure the variance in the date to assess the stability of the inference (see Materials and Methods).

(b) *ROLLOFF* analysis to test the effect of errors in the genetic map

To test the dependence of the *ROLLOFF* estimates on the precision of the genetic map, we systematically change the genetic map by modeling the change based on the convolution property of a gamma distribution with rate parameter ϕ. A low value of ϕ implies significant changes to the map and a high value allows for fine scale changes.

Suppose we model the genetic distance (d_i) between any pair of SNPs (s_1 and s_2) as a gamma distribution with two parameters such that k is the shape parameter and μ is scale parameter that is the reciprocal of the rate parameter (μ = 1/ϕ), then the probability distribution of distance d_i is given by

$$f(x; k, \mu) = \frac{x^{k-1} e^{-\frac{x}{\mu}}}{\mu^k \Gamma(k)} \quad \text{for x, } \mu \geq 0 \qquad (1)$$

Just by changing the rate parameter (ϕ), one can make fine scale changes to the map without changing the overall mean of the distribution. Using this method, we created different genetic maps by setting ϕ to range of values varying from 0.001 to 10 (multiples of 10). We simulated 10 individuals of mixed European and African ancestry that had 20% European and 80% African ancestry with the mixture occurring 10 and 100 generations ago. We used a non-overlapping dataset of European Americans and Nigerians as the reference populations and the inaccurate genetic map for modeling in *ROLLOFF*.

We observe that *ROLLOFF* is robust to errors in the genetic map and the estimated date falls within two standard deviations of the truth for the different maps (Table A.7).

(c) *ROLLOFF* analysis to test the effect of different bin sizes

To be computationally efficient, we divide the genome into bins separated by distance *d* and estimate the correlation between all possible pairs of SNPs in each bin. To test if the bin size has an effect on the results of *ROLLOFF*, we performed simulations with variable bin sizes and estimated the date of admixture. We simulated individuals of mixed European and African ancestry that had 20% European and 80% African ancestry with the mixture occurring 10 generations ago (10 individuals) and mixture occurring 100 generations ago (10 individuals). We performed four separate *ROLLOFF* analyses using a non-overlapping dataset of European American and Nigerians as reference populations with the bin size within the range of 0.001 - 1cM (we picked 5 values in this range).

For mixture occurring 10 generations ago, we observe that all the estimated dates fall within two standard deviations of the true (simulated) time depth. However, bin sizes of 0.5cM and greater contain very few points within each bin and thus the results should be considered with caution. We repeated the analysis for mixture occurring 100 generations ago and observed similar results. Based on these results, we conclude that the optimal bin size should be within the range 0.001 – 0.5 cM. For all our analyses, we use a bin size of 0.1cM (Table A.7).

(d) Simulation searching for bias in *ROLLOFF*

In many cases, genetic data for the true ancestral populations for a particular admixture event is not available either because the populations involved in the mixture are no longer extant (as the modern day population is very diverged from the ancestral population) or they have not been genotyped. In such cases, to the best that we can do is to use present-day populations that most closely resemble the true ancestral populations. To test the effect of using an inaccurate reference population on the estimation of a date by *ROLLOFF*, we constructed 10 diploid genomes of individuals of mixed ancestry with 20% European and 80% African ancestry that were simulated using CEU and YRI individuals but modeled using other reference populations. We performed these simulations for mixture occurring 10 generations ago as well as 100 generations ago. We performed *ROLLOFF* analysis with 4 sets of reference populations. These were:

(i) French Basque and Senegal Mandenka

(ii) Druze and Kenyan Bantu

(iii) HapMap3 Gujarati (GIH) and Kenyan Maasai (MKK)

(iv) Druze and Yoruba

In all cases, we observe that *ROLLOFF* can accurately estimate the date of mixture for the recent admixture date of 10 generations, even though inaccurate reference populations are used. Even at more ancient dates, the results are within two standard deviations of the true simulated date (Table A.8). This shows that *ROLLOFF* is robust to somewhat inaccurate parental populations and should be useful even in cases where it is difficult to obtain genetic data

from the true ancestral populations. In particular, it suggests that our method is likely to give unbiased results for West Eurasians regardless of the sub-Saharan African ancestral population chosen.

To assess if inaccurate ancestral populations give unbiased results for a real admixed population, we applied *ROLLOFF* to data from African Americans (HapMap3 ASW), using Senegal Mandenka and Basque as reference population. We estimated that the date of admixture for ASW is 6 ± 1 generation, which is consistent with previously published reports [5; 7] and with the date estimated if YRI and CEU (which are likely close to the true ancestral populations) are used as reference populations. The date remains unchanged even if Druze and YRI are used as reference populations.

In addition, we carried out simulations to test the performance of *ROLLOFF* in situations of very low mixture proportions and old mixture dates as seen in Southern Europeans and Jewish groups. We simulated data for 20 individuals of mixed YRI and CEU ancestry with the mixture proportion and date of mixture selected to match Southern Europeans (1-3% mixture proportion, 55 generations ago) and Jewish groups (3-5% mixture proportion, 89 generations ago). We then ran *ROLLOFF* with the same set of inaccurate reference populations as shown above. We repeated each simulation 100 times and then computed the average date and bias (here, *bias* is defined as (average-truth)/(truth)).

We observed that in the case of low mixture proportions and old mixture dates, there is an upward bias in the estimated dates (Table A.9). This effect is attenuated as the number of admixed samples increases (Table A.10) and as the mixture proportion increases (Table A.11)

but does not seem to be significantly affected by the ancestral populations used as reference in the *ROLLOFF* analysis (Table A.9).

To test how much this effect is biasing our estimated dates for West Eurasians, we performed simulations to generate data for individuals of mixed European and African ancestry where we set the mixture proportion (θ), time since mixture (λ) and number of samples to match the parameters estimated for each West Eurasian group individually. We then performed *ROLLOFF* analysis using HapMap3 Italian Toscanis (TSI) and Kenyan Luhya (LWK) as reference populations. We repeated each simulation 100 times and computed the average and bias. The bias is in general very small except in the case where the mixture dates are old and sample sizes are small, which is the case for the Druze and most of the Jewish groups, where the bias is typically at least 20% (Table 2.2 and Table A.12). We discuss this bias and report a bias correction in the main text.

(e) *ROLLOFF* analysis for double admixture scenarios

A potential pitfall in estimating dates of admixture is that the historical mixture event may not have occurred all at once, but instead may have taken place over multiple different times (pulse migration model) so that the pattern in the data in fact reflects a range of mixture times. To explore this, we ran *ROLLOFF* to infer the date of admixture on data simulated under a double admixture scenario (where there were two distinct events of gene flow between the populations). We simulated double admixture scenarios in which a 90%/10% admixture of CEU and YRI occurred at λ= 30, followed by a 60%/40% mixture of that admixed population and YRI at λ= 6. We also simulated data for a 50%/50% admixture of CEU and YRI that occurred at

λ= 30, followed by a 50%/50% mixture of that admixed population and YRI at λ= 10 and 80%/20% admixture of CEU and YRI at λ= 40, followed by a 20%/80% mixture of that admixed population and YRI at λ= 20. For each simulation, genomes of 20 admixed individuals were constructed and *ROLLOFF* analyses were performed using a non-overlapping dataset of European Americans and Nigerians as reference populations.

Applying *ROLLOFF* to the simulated data, we observed that for case 1 where the mixture occurred at λ = 30 generations followed by λ= 6 generations, the date of admixture was estimated at 34 and 5 generations, when we fitted a sum of two exponentials and 6 generations when we fitted a single exponential distribution to the decay of the correlation coefficients. Similarly, *ROLLOFF* was able to estimate the date of admixture for the simulations for double admixture of 50%/50% mixture of CEU and YRI that occurred at λ= 30, followed by a 50%/50% mixture of that admixed population and YRI at λ= 10 accurately (35 and 9 generations when output was fitted with a sum of two exponentials and 11 generations when fitted with a single exponential) (Figure A.8). However, for simulations of more ancient mixture dates of 80%/20% admixture of CEU and YRI that occurred at λ = 40, followed by a 20%/80% mixture of that admixed population and YRI at λ = 20, we could only reliably estimate the date for the more recent admixture event (24 and 2 generations when output was fitted with a sum of two exponentials and 22 generations when fitted with a single exponential). Standard errors were not computed for this analysis, as it is not clear how to apply the standard Jackknife theory for analysis with a sum of exponentials.

The dates for the recent admixture event were qualitatively similar to the true time depth, when the data was fitted with a single exponential or sum of two exponentials. A caveat is that

we have only simulated a limited number of double mixture scenarios. In principle, further exploration of different values of (θ, λ) might identify situations in which we could estimate the date of the older mixture event using *ROLLOFF*.

(f) Simulations of continuous admixture

To model continuous gene flow, we simulated recurrent mixture over a specified number of discrete generations- varying between 1 to 100 generations (graduation mixture model). In each simulation, we generated data for individuals of mixed ancestry using two ancestral populations (CEU and YRI), where the gene flow occurred in an interval $I = [a,b]$ where $0 \leq a \leq b$. In each generation during I, we allow a proportion m (computed based on mixture proportion (θ)) of YRI lineages to migrate, yielding a total of 20% average African ancestry in the resulting admixed samples (Figure A.9). At generation 1, we sample African haplotypes with probability (θ) and European haplotypes with probability (1-θ). We resample ancestry at each marker with probability $1\text{-}e^{-g}$, where g is the genetic distance between markers (in Morgans). Once the ancestry is sampled, a haplotype is copied from an individual of the chosen population (YRI or CEU) and copied to the genome of the admixed individual and the process is continued until the end of chromosome is reached. After the first generation, we allow there to be uni-directional migration from YRI into the admixed individuals. The pool of ancestral haplotypes is updated in each generation with haplotypes from the previous generation and this procedure is repeated until the end of the interval I. Next, if parameter $a \neq 0$, then following YRI mixture, there are a generations of random mixture between the admixed individuals only. This can roughly be thought of as simulating genetic drift, since admixture. This procedure is repeated to create the

genomes of 20 admixed individuals and pairs of haploid individuals are combined to construct 10 diploid admixed individuals.

In order to test the performance of *ROLLOFF,* we performed 30 simulations. In each simulation, we varied the values of the length of the interval $I = b\text{-}a$ and the time since mixture *(a)*. We performed *ROLLOFF* analysis using a non-overlapping dataset of 1,107 European American and 737 Nigerian Yoruba individuals as reference samples. All analyses were restricted to 339,171 SNPs and the fine scale recombination map by Myers et al. [6] was used for mapping the genetic distance.

We applied *ROLLOFF* to the simulated data, fitting a single exponential decay in each case. We observe that when the interval *I* is very small, the *ROLLOFF* result correlates to the time since the last mixture event. However, as length of the interval *I* increases, the estimated dates reflect averages of mixture dates over a range of the interval (Table A.13).

(g) *ROLLOFF* analysis in cases of no mixture related to the reference populations

We performed *ROLLOFF* analysis to estimate the date of admixture in the East Asian Uygur (HGDP-CEPH- Uygur) population who are known to have both West Eurasian and East Asian ancestries [8; 9]. We used extremely inaccurate and unrelated populations— African Pygmies (HGDP-CEPH- Mbuti and Biaka Pygmies) and Nigerian YRI—as reference populations to test the performance of the *ROLLOFF* in cases when there is no admixture related to the reference populations. To contrast the situation of no mixture and to ensure that there are no technical issues with the dataset, we simulated 20 individuals of mixed Pygmy and Yoruba ancestry as positive control. These samples were simulated using HGDP-CEPH Biaka and Mbuti Pygmies

and HGDP-CEPH Yoruba individuals as ancestral populations with 20% Pygmy ancestry and 80% Yoruba ancestry with the mixture occurring 10 generations (10 individuals) and 100 generations ago (10 individuals). We used a dataset containing 591,320 SNPs and used Pygmies and YRI as reference populations for *ROLLOFF*.

We observed clear evidence of admixture in the simulated individuals as we see an approximately exponential decay of LD with distance in the simulated individuals, with estimated dates of mixture as 10 and 90 generations. However, we observe that the correlation is almost zero in the Uygur population (Figure A.11). This is consistent with expectation as these populations do not have Pygmy and YRI ancestry or ancestry from populations closely related to Pygmy or YRI.

Text A.5: Searching for the source of African ancestry in West Eurasians

In order to identify the source of the African ancestry in Levantine, Southern Europeans and Jewish groups, we performed Principal Component Analysis (PCA) and *4 Population Test*[3]. We first started by establishing axes of variation in Africa by performing - (a) PCA with 15 sub-Saharan African populations and (b) PCA with sub-Saharan Africans, South Africans (HGDP-CEPH- San) and European Americans (HapMap3- CEU). The goal of these analysis was to investigate if we can reliably distinguish ancestry from various parts of Africa and to ensure that none of the African populations have any West Eurasian ancestry. If we include samples that have West Eurasian ancestry, the PCA will be biased toward the population more closely matching the West Eurasian group related to the mixture, which would confound our results. For these analysis, we used the dataset of 10 West and South African populations from Bryc et al. [10],

three sub-Saharan African populations from HGDP-CEPH [9] and five populations from HapMap3[11].

The PCA in Figure A.12 shows that most of our samples fall along two main axes of variation, which we call Chadic (e.g. Bulala, Mada and Kaba) vs Non- Chadic and East Africans (e.g. Kenyan Luhya- LWK) vs West Africans (e.g. Nigerian Yoruba - YRI). This pattern is similar to one previously observed by Bryc et al. [10]. In addition, we identified that the East African Maasai have some West Eurasian ancestry as samples vary in their proximity to CEU in the PCA, and hence we did not include them in subsequent PCA explorations.

To make a qualitative inference about the source of the African ancestry in West Eurasians—relative to the Chadic and East-West axes of variation in Africa defined in the PCA analyses—we performed PCA Projection. Specifically, we carried out PCA Projection analysis with three sub-Saharan African populations that are at the extremes of the PCA in Figure A.12: Bulala, Yoruba, Luhya that we take to represent Chadic, Niger-Kordofanian, and Nilo-Saharan related ancestry respectively. We also used Asians (HapMap3- CHB) to represent non-African variation. The value of using the CHB rather than a West Eurasian population here is that the CHB are likely to be symmetrically related to all West Eurasians. Hence, including them in the analysis will not bias the results toward matching one West Eurasian group more than another.

We performed projection PCA with all three possible pairs of African populations along with CHB, and then plotted the mean values of all the samples from each West Eurasian population onto the first and second PCs. As a reference, we also project African Americans (HapMap3 ASW) and North Africans (HGDP-CEPH- Mozabite) that have inherited a mixture of both sub-Saharan African and West Eurasian ancestry, as well as Northern Europeans (CEU).

Figure A.13 shows that West Eurasians without any evidence of sub-Saharan African ancestry (like CEU) all fall on a single point on the plot, as expected. However, the West Eurasians with sub-Saharan African ancestry fall along a gradient pointing toward some sub-Saharan African populations more than others. The Figure A.13A and Figure A.13B show that the African ancestry in West Eurasians is likely not related to Chadic Bulala population as in both cases the West Eurasians are pointing away from Bulala. However, when we perform the analysis using Luhya, Yoruba and CHB to construct the PCs, we observe that the West Eurasians are pointing to a population that is intermediate between Kenyans and Yorubans, but somewhat more closely related to the East Africans.

In order to formally test if Levantine, Southern Europeans and Jews are more closely related to Luhya (compared to Yoruba), we perform the *4 Population Test* with the tree ((LWK, YRI),(*X*, CEU) that is consistent with the data, where *X* is a West Eurasian populations (Table A.14). We are not able to reject this tree when *X* is any of the Southern European or Jewish groups, and hence we cannot formally reject the hypothesis of at least some West African ancestry in these groups. However, we are able to show that the African ancestry in a couple of Levantine populations is more closely related to East Africans than West Africans. It is historically plausible that gene flow occurred from both West Africa (there were slave caravans across the western Sahara in Roman times[12]) and from East Africa, via the Egypt and Middle East[13].

It is important to note that the methods used in our study for inferring ancestry proportion as well as the date of admixture are carefully designed so that they are not sensitive to which African population is used, as long as the phylogeny is correct. Thus, our inability to pinpoint the African source population for the sub-Saharan African ancestry in West Eurasians is not

expected to bias these inferences. To confirm this expectation, we performed our analysis with LWK (instead of YRI) to represent the sub-Saharan African source population and show that our results remain qualitatively similar. Results for f_4 *Ancestry Estimation* and *ROLLOFF* using East Africans and CEU as ancestral populations are shown in Table A.15.

A. Northwest Europe

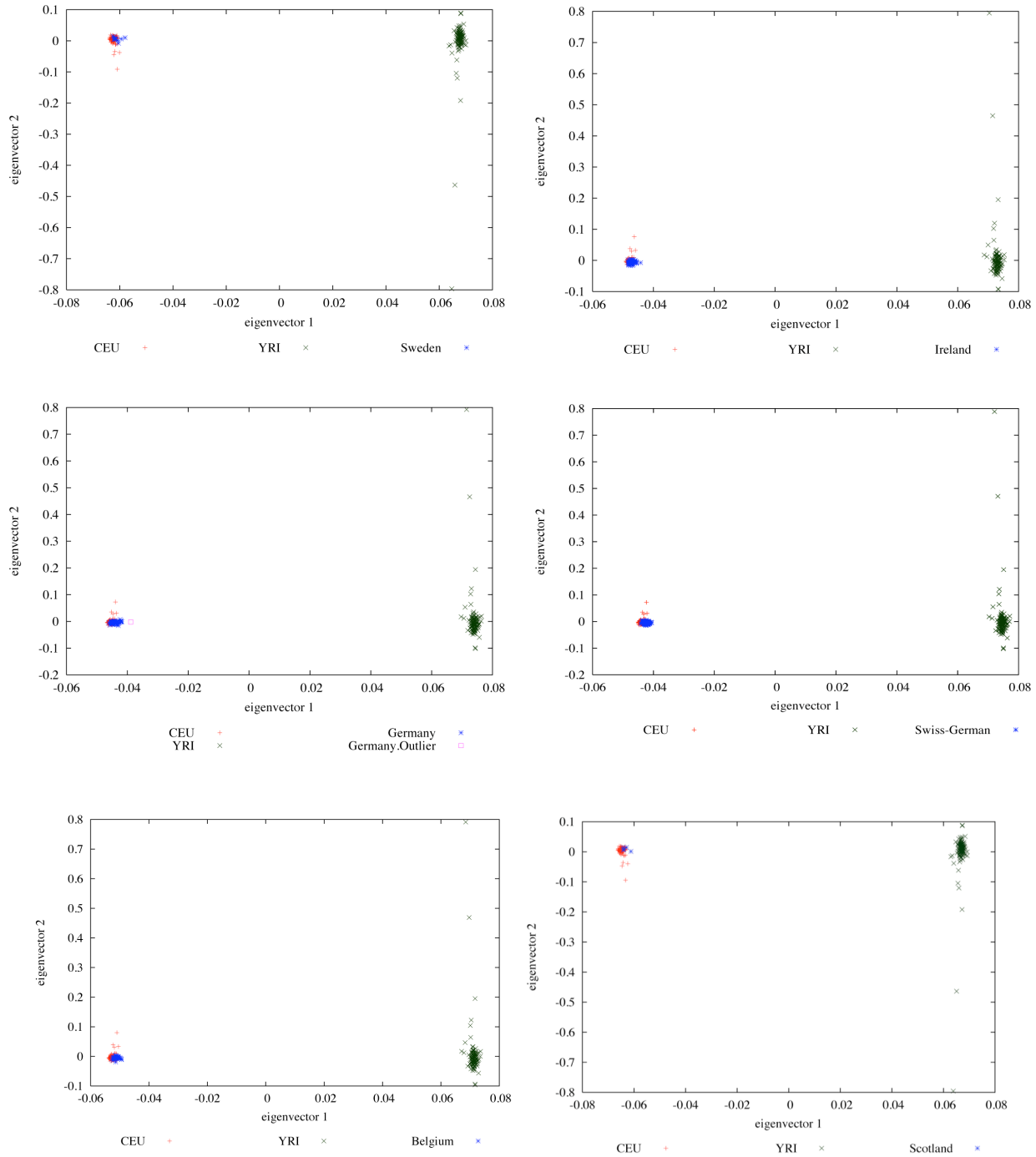


Figure A.1: PCA-based search for outliers and substructure in West Eurasians.

Figure A.1 (Continued)

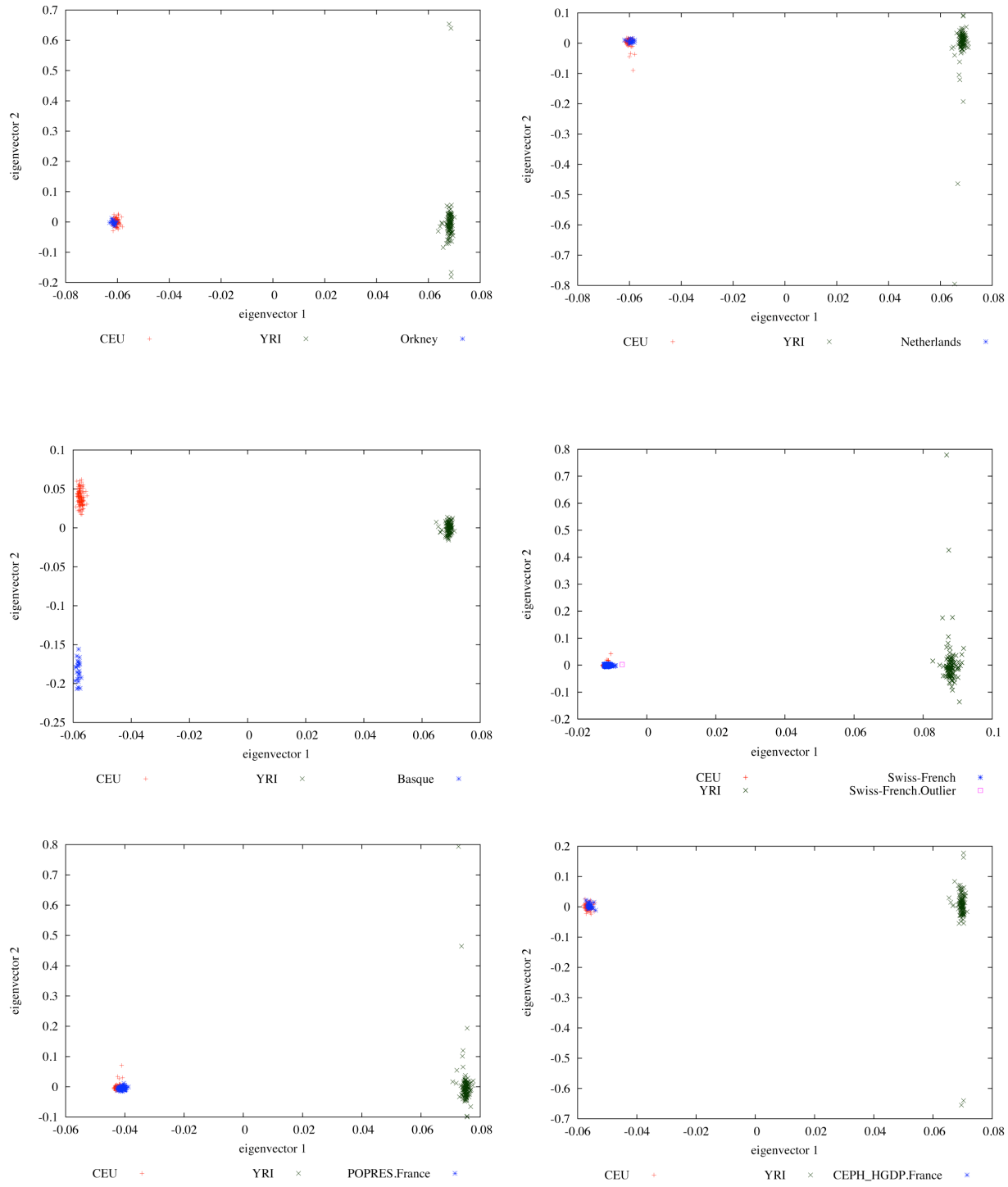


Figure A.1 (Continued)

B. East-Central Europe

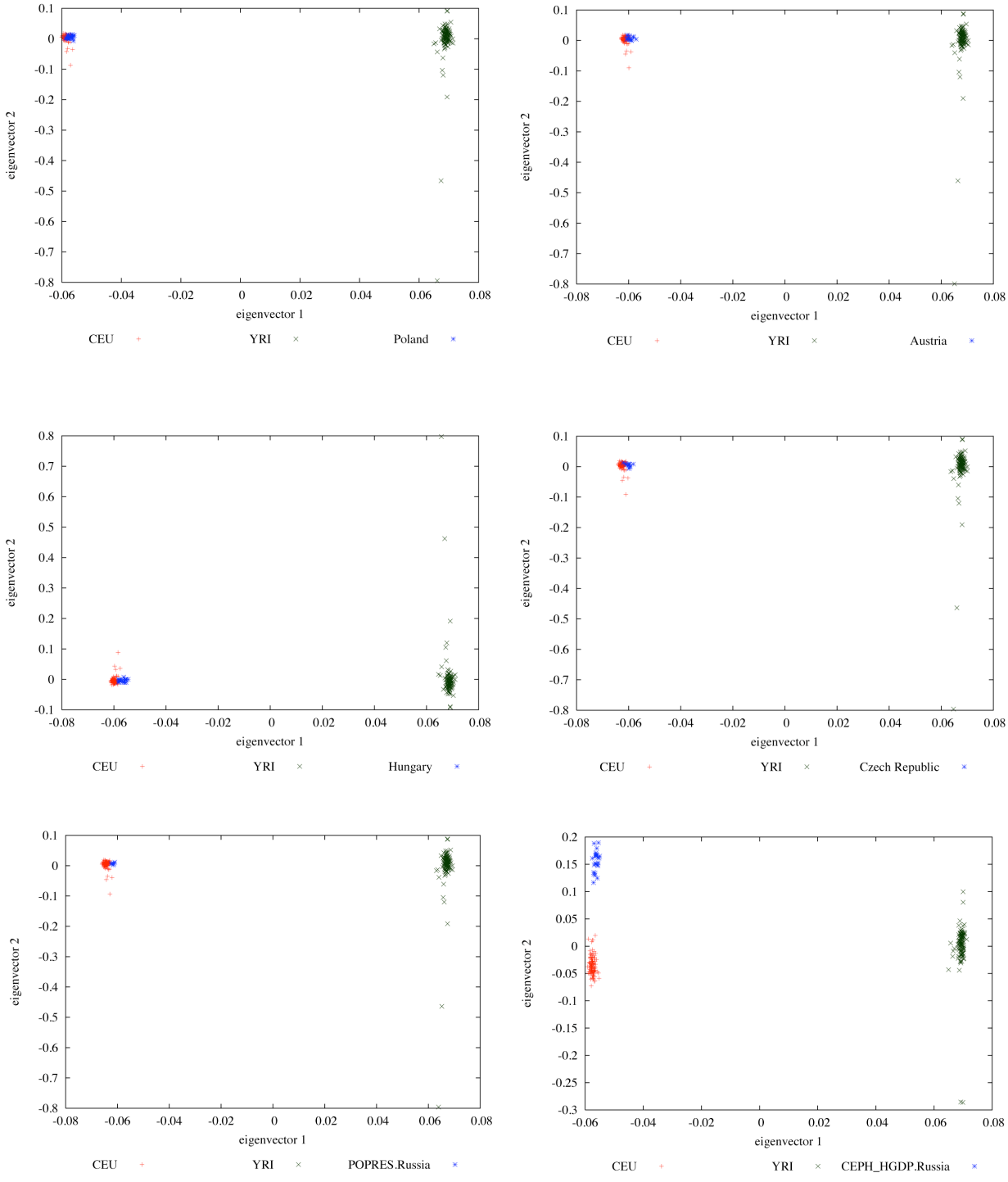
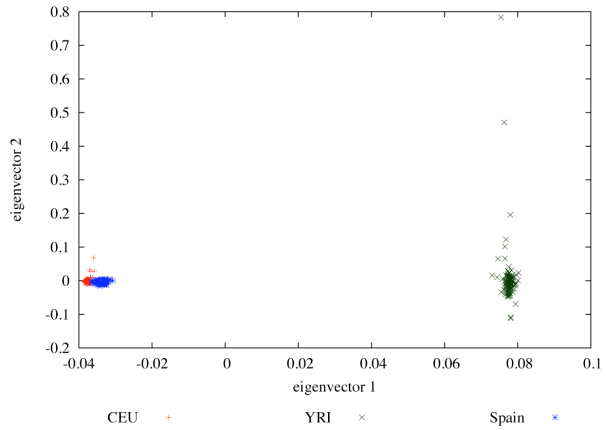
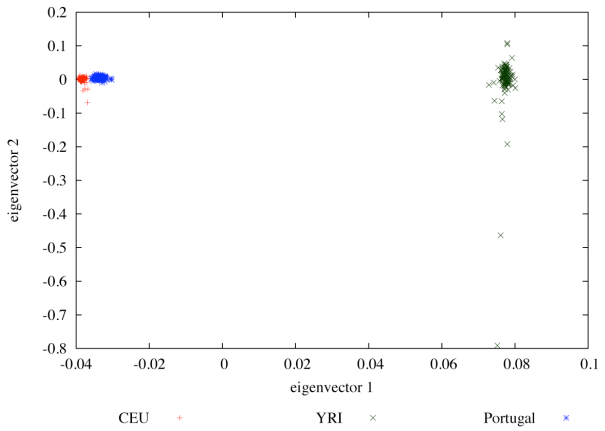
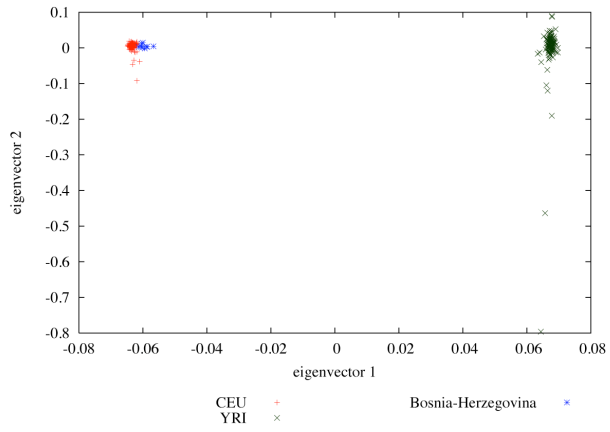


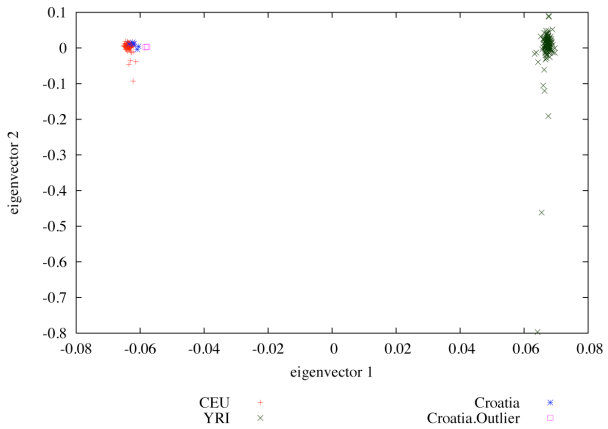
Figure A.1 (Continued)

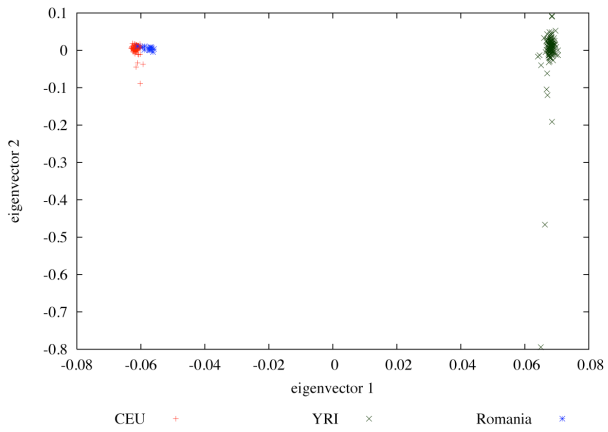
C. Southern Europe











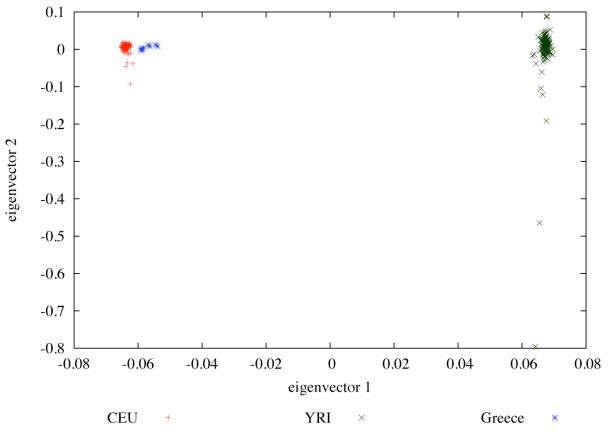
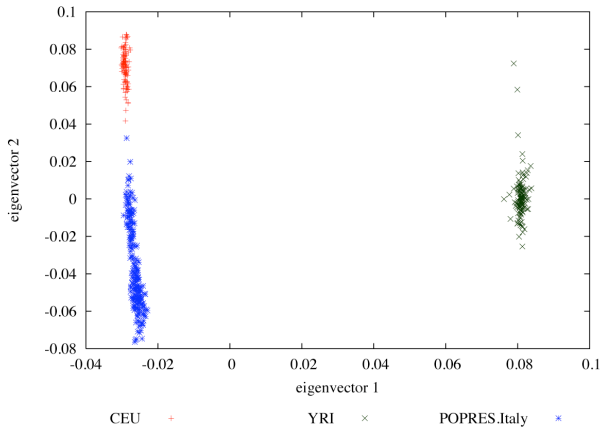
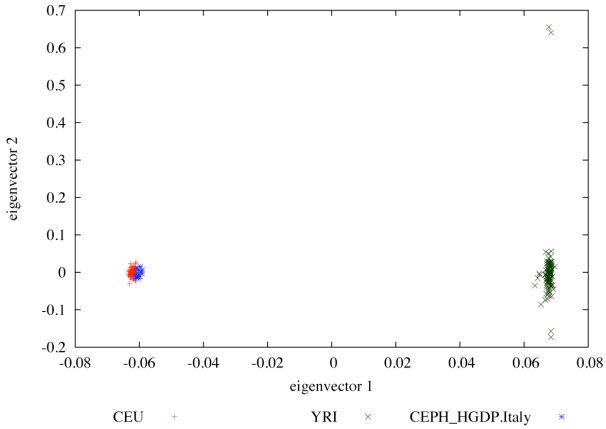
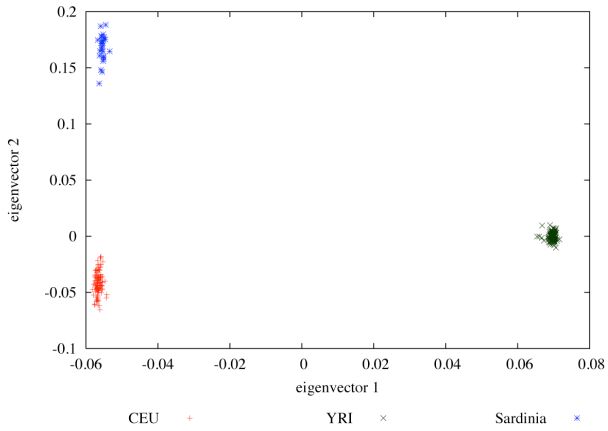
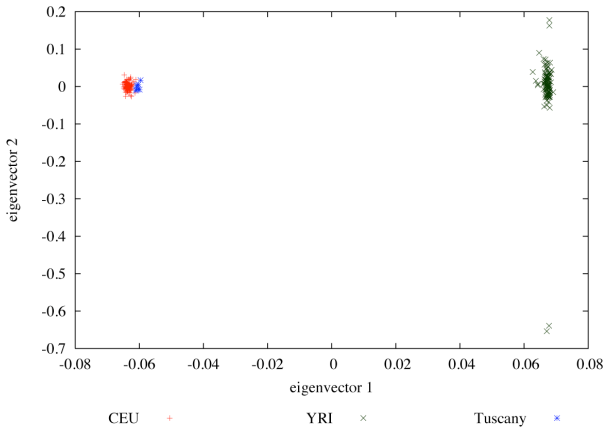


Figure A.1 (Continued)









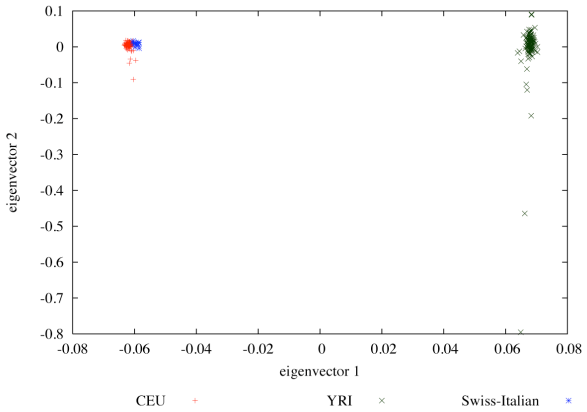
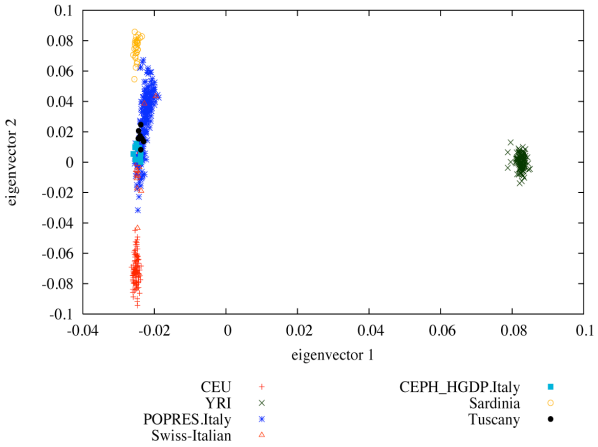
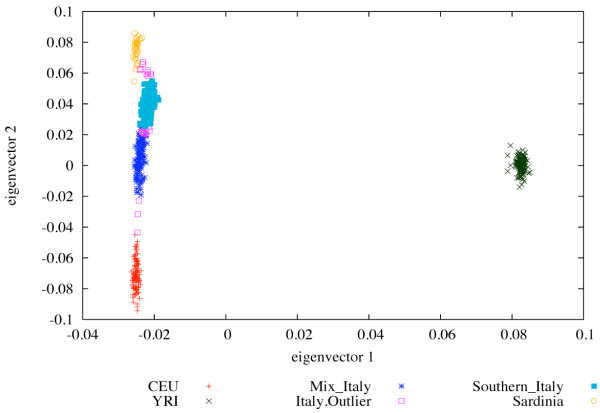


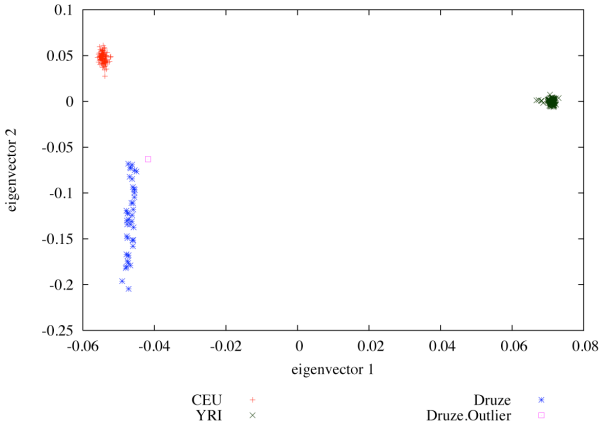
Figure A.1 (Continued) Combined PCA analysis for all Italian populations:

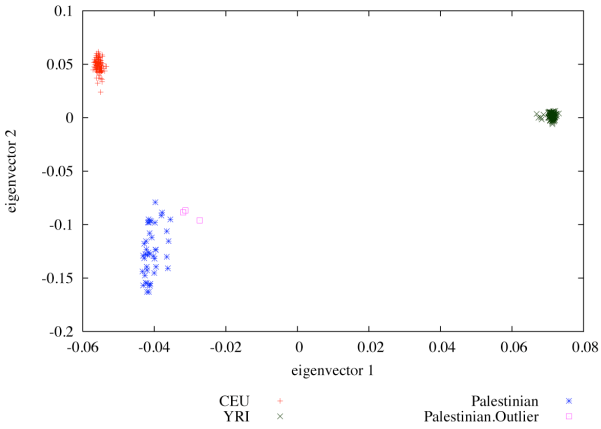
a. Before outlier removal **b. After outlier removal and relabeling**





D. Levant





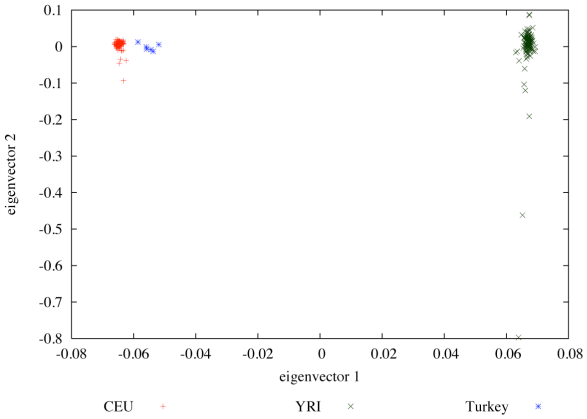
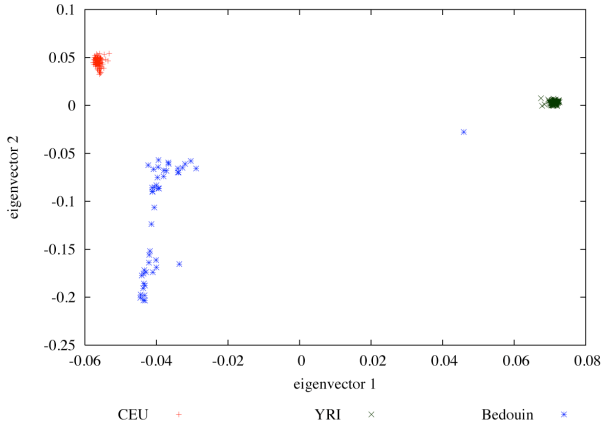


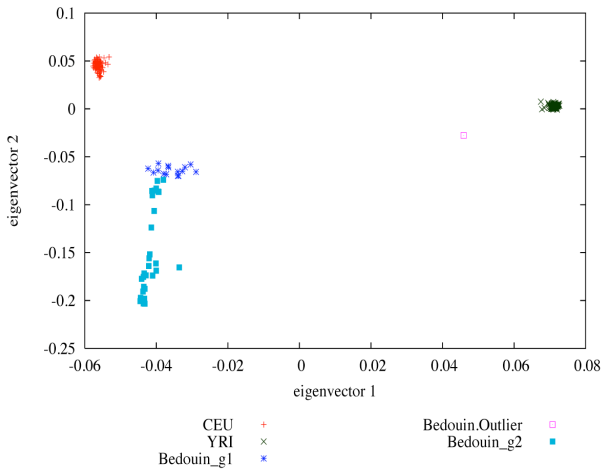
Figure A.1 (Continued)

Bedouin

a. Before outlier removal



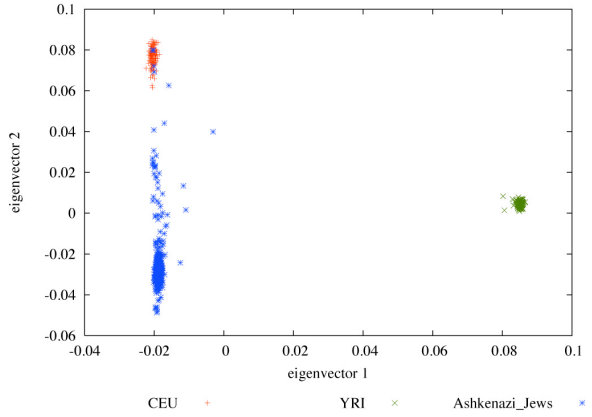
b. PCA-based classification



E. Jewish Groups

IBD.Ashkenazi Jews

a. Before outlier removal



b. PCA-based classification

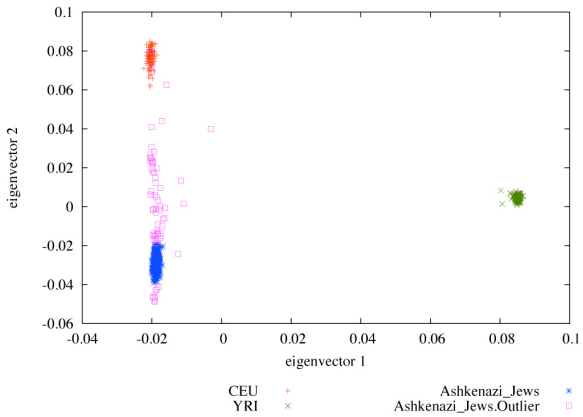


Figure A.1 (Continued)

Jewish HapMap Project

Combined PCA analysis for all Jewish groups:

a. Before outlier removal **b. After outlier removal**

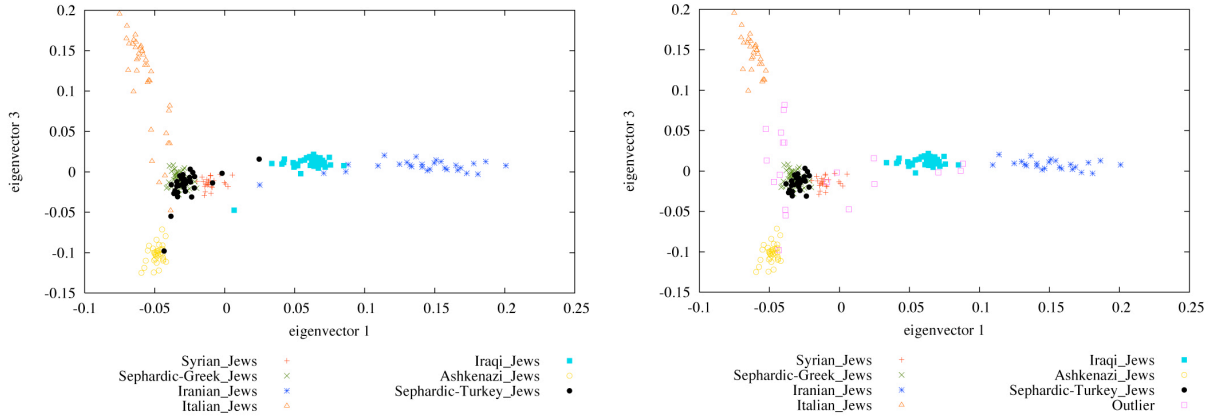


Figure A.1 (Continued) PCA was performed using YRI, CEU and *X* (where *X* = any West Eurasian population). A plot of the first and second PCs is shown all West Eurasian populations. Outliers (if any) are shown in pink boxes and labeled as *X*.Outlier. In three populations – Bedouins, Italians and Ashkenazi Jews, we observe significant population structure. The populations have been divided into multiple groups and PCA results both before outlier removal and reclassification are shown below.

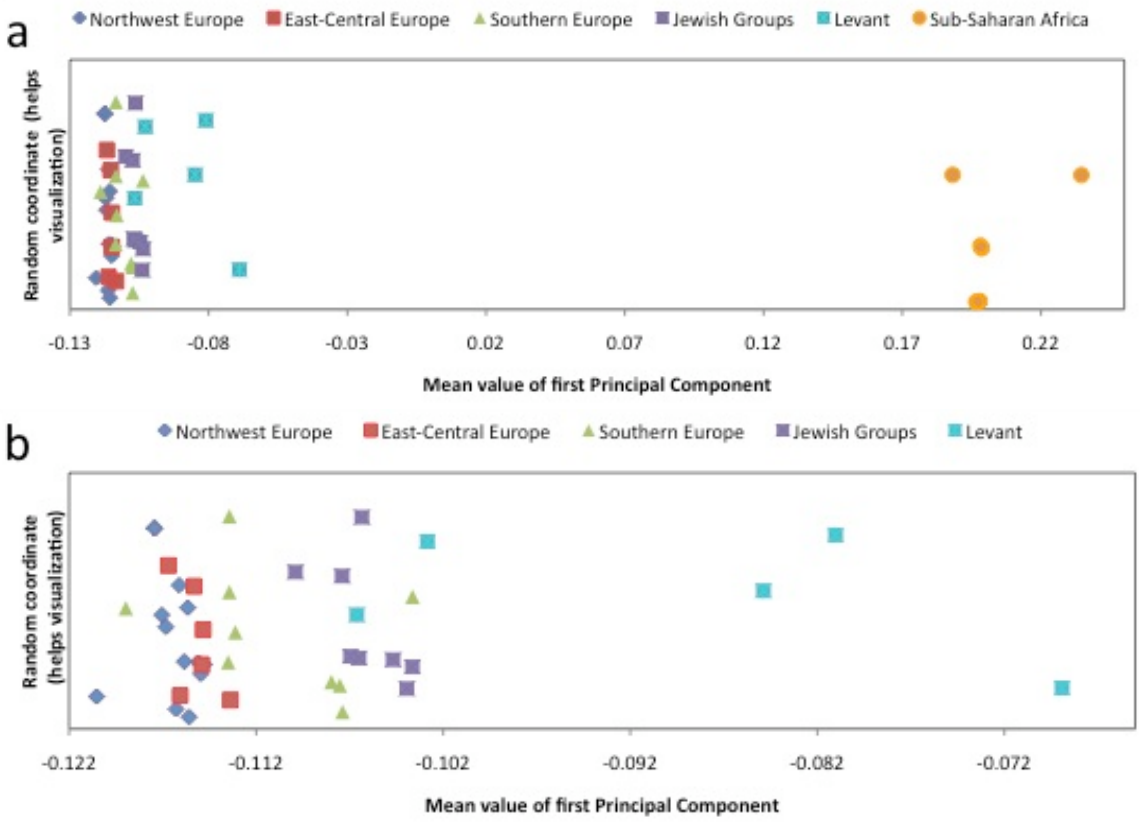
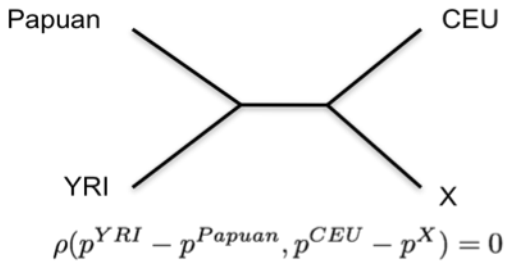


Figure A.2: PCA Projection with Adygei and Kenyan Bantu. PCA was performed using genome-wide SNP data from Adygei and Kenyan Bantu. All West Eurasians populations with samples sizes of ≥5 were then projected onto these PCs. (a) The first panel presents data for all populations, (b) while the second provides a higher resolution view of West Eurasians after removing Sub-Saharan Africans. Each point on this graph indicates the mean value of the first PC for a projected population and West Eurasians populations are colored by 5 regional groupings—"Northwest Europe", "East-Central Europe", "Southern Europe", "Levant", "Jewish Groups"—with the assignments of populations to groups as shown in Table 2.1. The grouping "Sub-Saharan Africa" refers to six populations from the HGDP-CEPH panel: Kenyan Bantu, South African Bantu, Mandenka, Mbuti Pygmy, Biaka Pygmy and Yoruba. A qualitatively similar pattern is seen as in Figure 2.1.

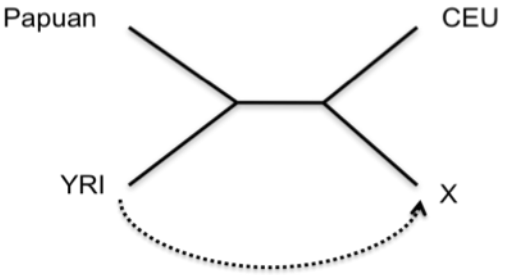
(A) 4 Population Test

Simple Topology



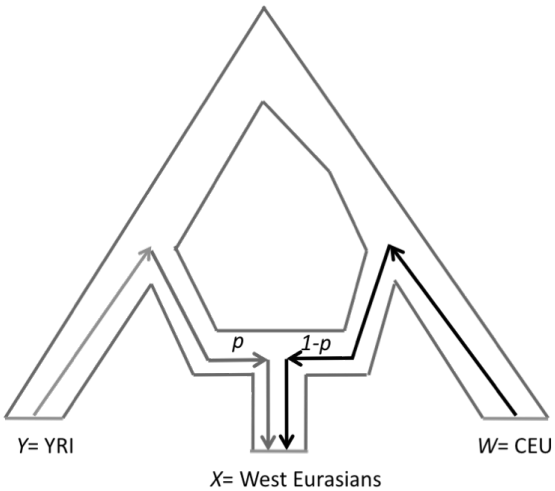
$$\rho(p^{YRI} - p^{Papuan}, p^{CEU} - p^{X}) = 0$$

Tree topology with mixture



$$\rho(p^{YRI} - p^{Papuan}, p^{CEU} - p^{X}) > 0$$

(B) 3 Population Test



$$f_3(X;Y,W) = \frac{\sum_{i=1}^{n}\left(p^{X}{}_i - p^{Y}{}_i\right)\left(p^{X}{}_i - p^{W}{}_i\right)}{\sum_{i=1}^{n} p^{X}{}_i\left(1 - p^{X}{}_i\right)}$$

Figure A.3: Formal Tests of admixture.

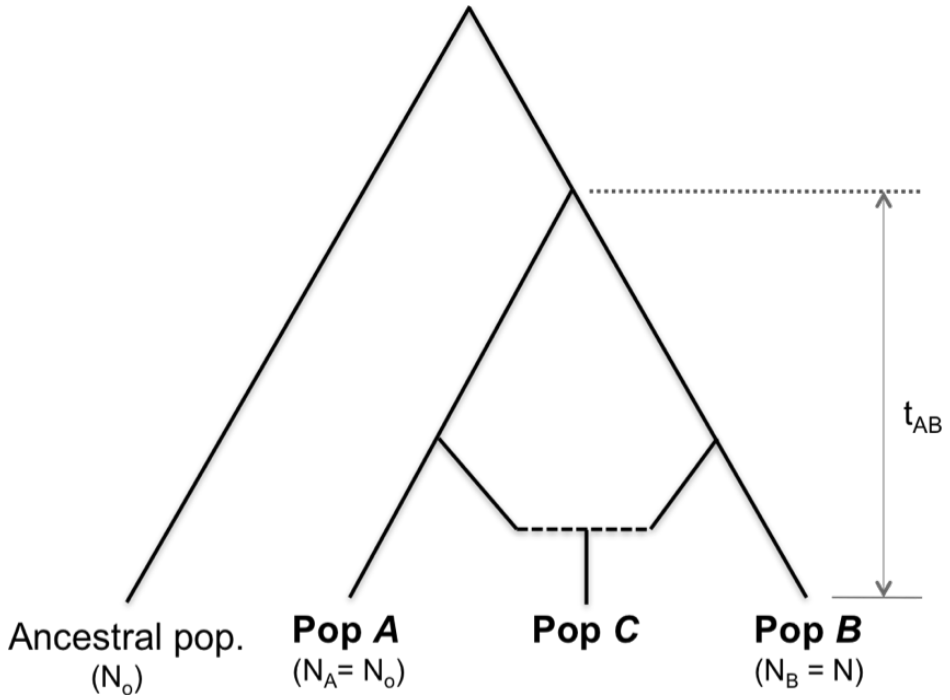


Figure A.4: Demographic model to test the effect of ascertainment bias on *3 Population Test*. Simulation framework was adapted from reference [[1]] where the authors were simulating data to test effect of SNP ascertainment bias on calculation of F-statistics similar to *3 Population Test* statistics.

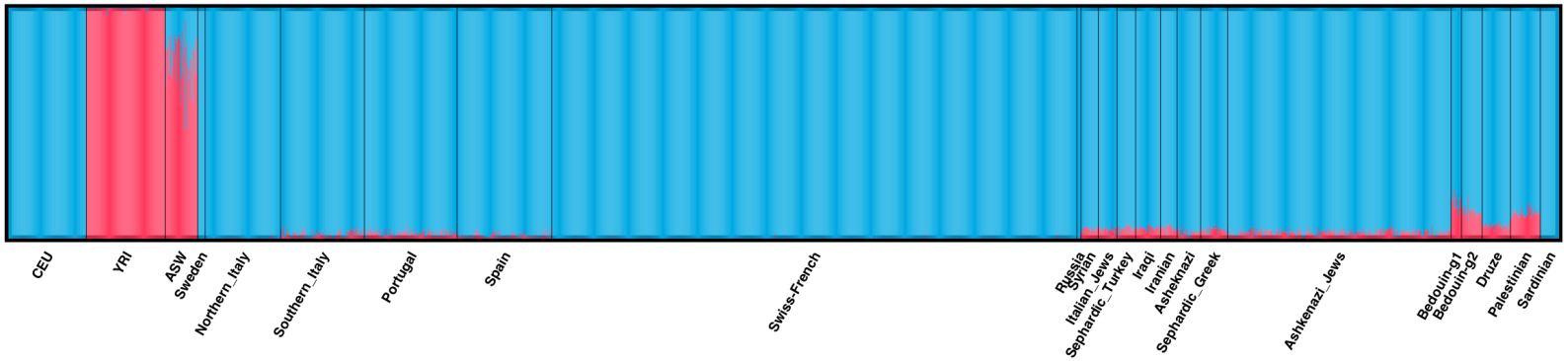


Figure A.5: Estimation of African ancestry using STRUCTURE. We applied STRUCTURE 2.2 to estimate the mixture proportions using ~13,900 markers (selected to not be in LD with each other) and K=2. Each individual is represented by a single line with the length of the different colors reflecting the individual ancestry proportions.

Figure A.6: Geographic gradient of African ancestry in Europeans. Sub-Saharan African ancestry proportions were estimated using *f_4 Ancestry Estimation.* Populations in grey are estimated to have sub-Saharan African ancestry between 1-4%. The * in Switzerland indicates that the three populations available from this country have variable estimates: Swiss-Germans show no evidence of African mixture, Swiss-French 0.5 ± 0.2% and Swiss-Italians 1.6 ± 0.2%. The '+' sign in Italy indicates that multiple samples were available but all show evidence of African mixture. No data are available from countries filled with diagonal lines. The map was downloaded from- http://www.ecozon.com/images/europe_map.jpg

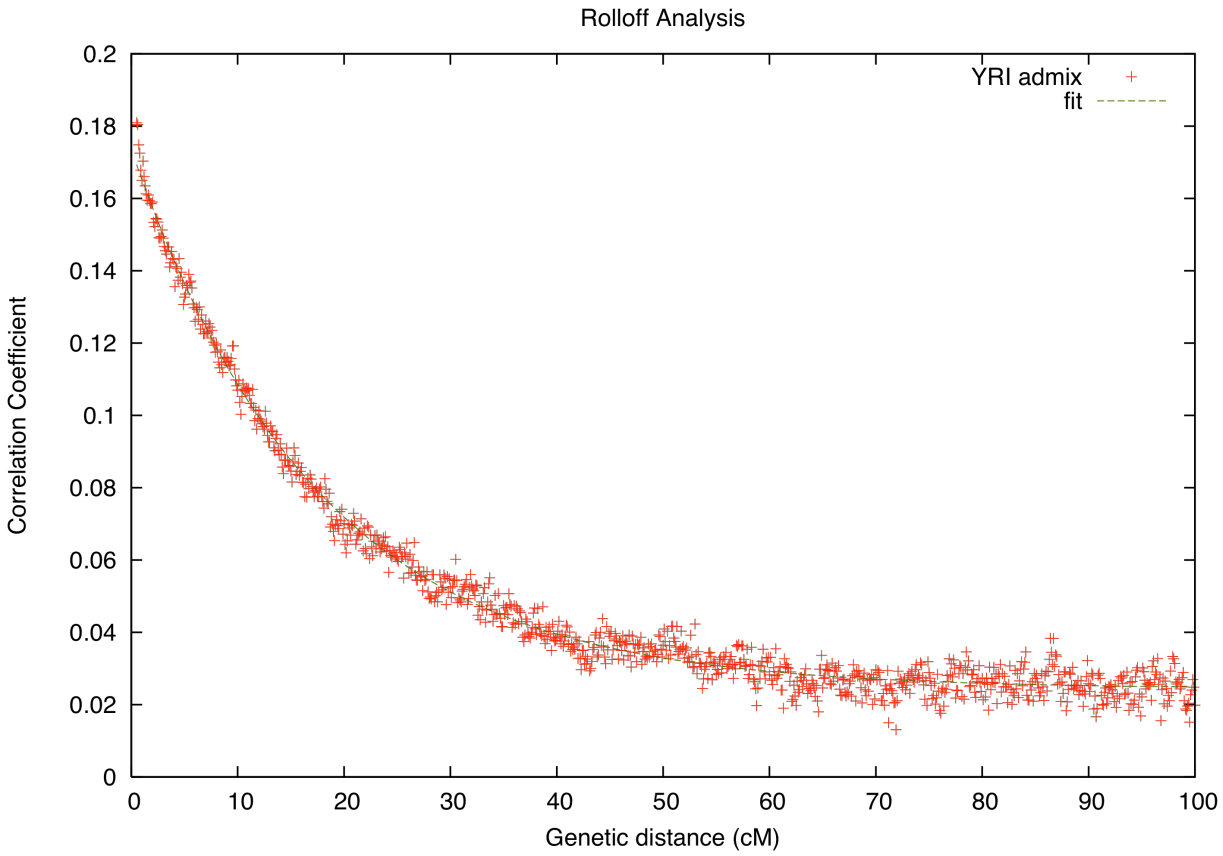


Figure A.7: *ROLLOFF* simulation for a scenario similar to African Americans. We constructed genomes of 10 individuals with mixed European and African ancestry. We set the time since mixture (λ) at 6 generations and the European ancestry proportion (θ) was sampled from a beta distribution with mean 20% and standard deviation 10%. We performed *ROLLOFF* analysis with a non-overlapping dataset of European Americans and Yoruba Nigerians as reference populations. We plot the decay of admixture LD as a function of genetic distance and estimate the date of admixture as 6 ± 1 generations, by fitting an exponential distribution to the data.

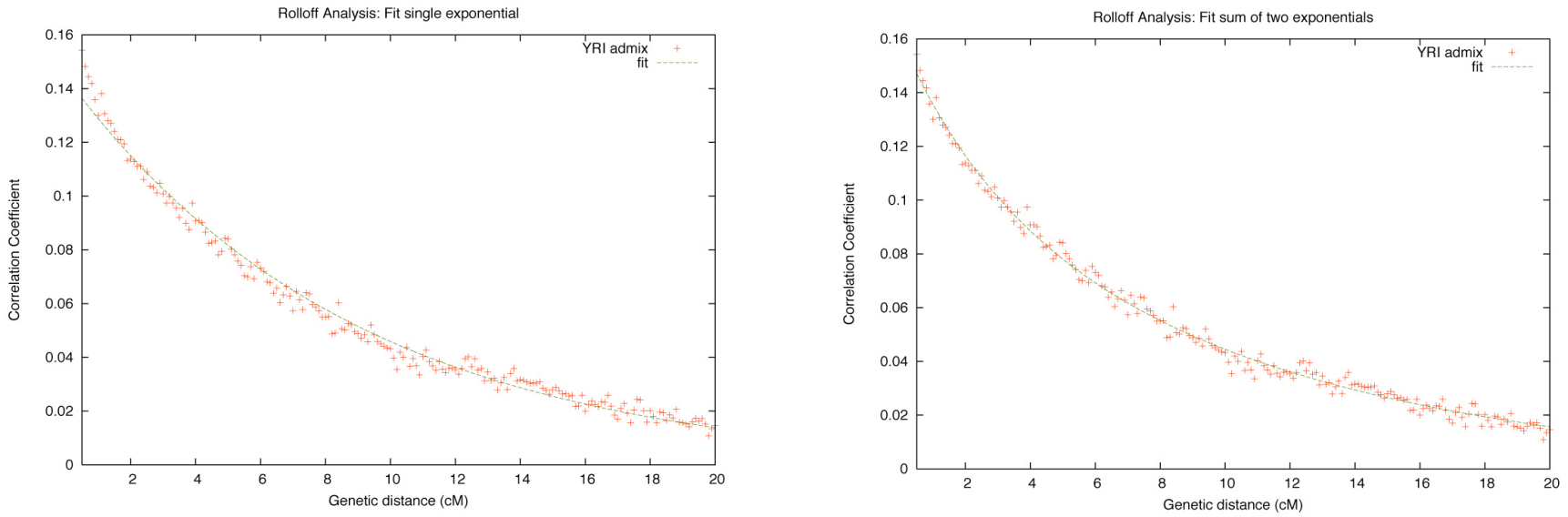


Figure A.8: *ROLLOFF* analysis for double admixture event. We simulated double admixture scenarios (two events of gene flow) in which a 50%/50% mixture of CEU and YRI that occurred at λ= 30, followed by a 50%/50% mixture of that admixed population and YRI at λ= 10 generations. We performed *ROLLOFF* analysis using a non-overlapping set of Yoruba Nigerians and European Americans as reference populations. In the left panel, we fit a single exponential distribution to the data and estimate the date of the admixture event as 11 generations. In the right panel, we fit a sum of two exponentials to the data and estimate the dates of admixture as 35 and 9 generations. In both cases, we accurately estimate the date of the most recent mixture event.

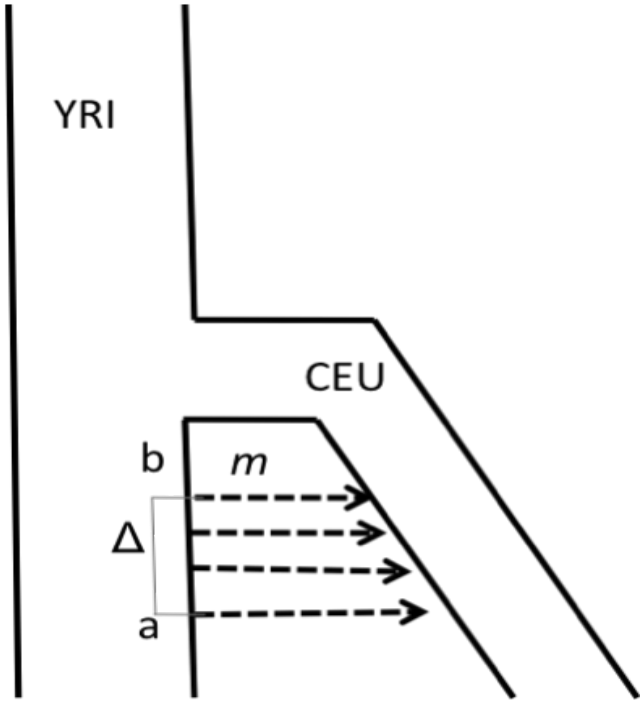
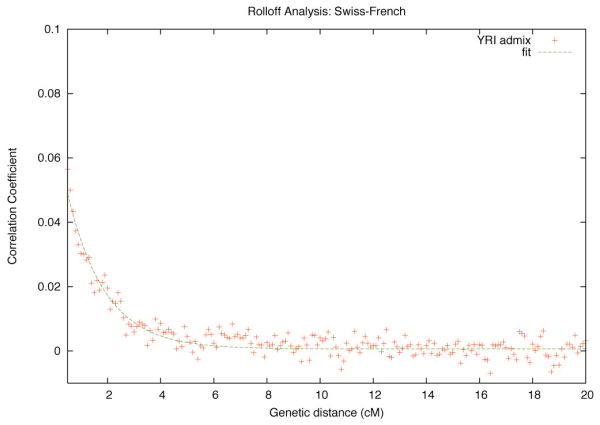
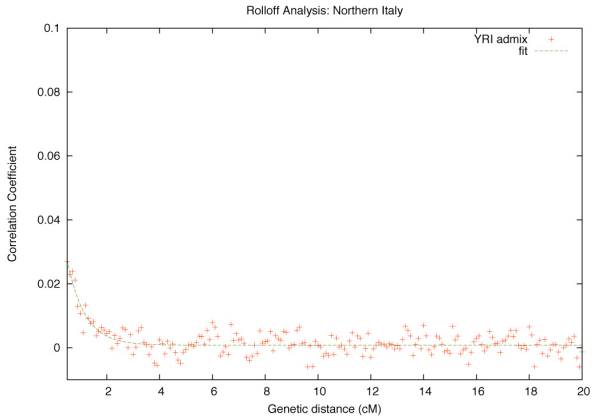


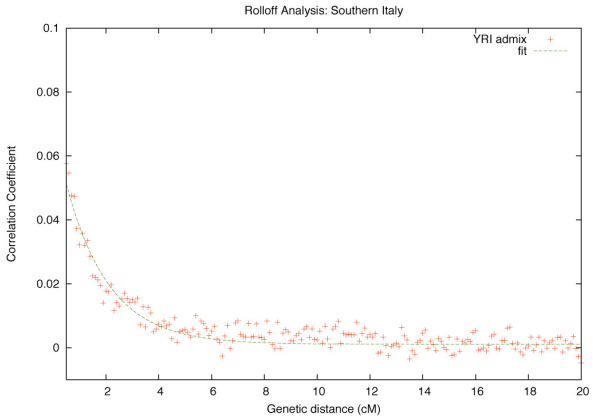
Figure A.9: A demographic model for continuous admixture scenarios. In order to test the performance of *ROLLOFF* under continuous admixture scenarios, we simulate data for individuals with mixed ancestry using data for two ancestral populations CEU and YRI, where the gene flow occurs in an interval $I = [a,b]$ where $0 \leq a \leq b$ and the time is in generations. In each generation, a proportion m of the population is replaced by migrants from the other. Based on this model, we simulate data for Population X where we set the overall ancestry of YRI = 20% and vary the values of a and b.

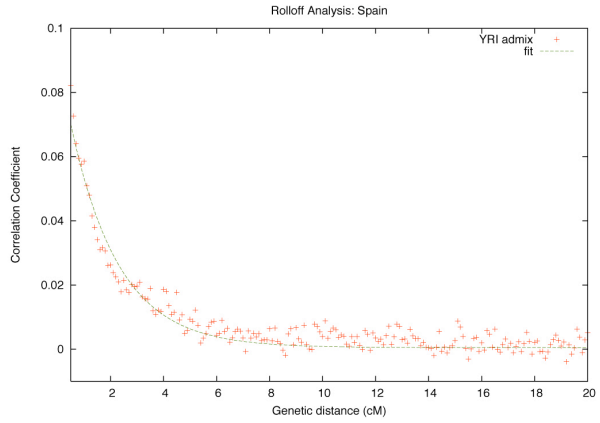
A. Northwest Europe



B. Southern Europe







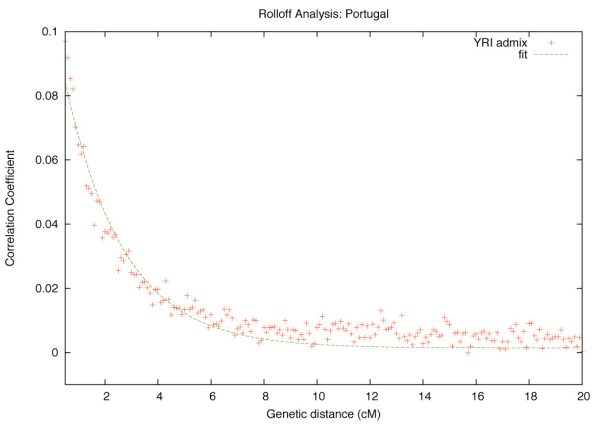
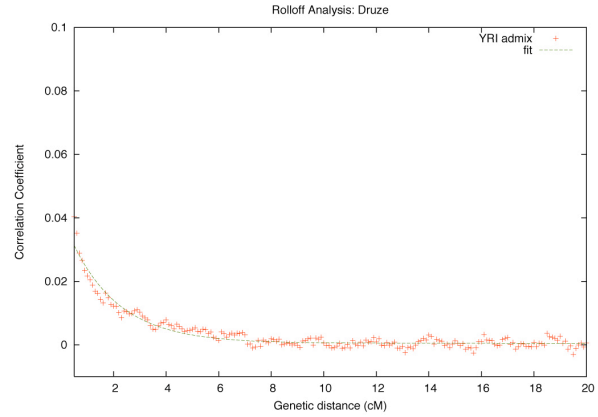
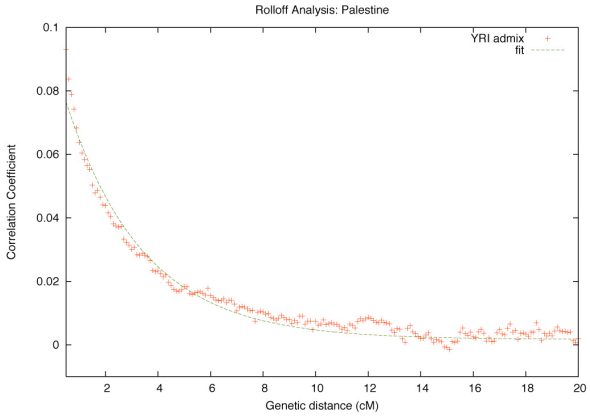


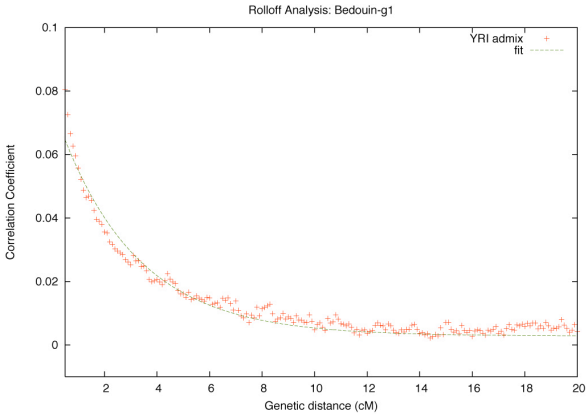
Figure A.10: *ROLLOFF* analysis for West Eurasians.

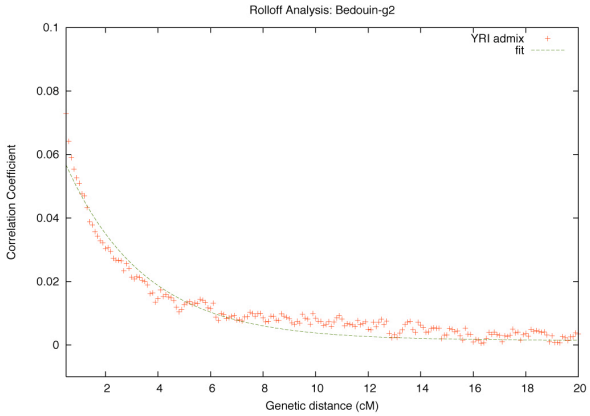
Figure A.10 (Continued)

C. Levantine Populations

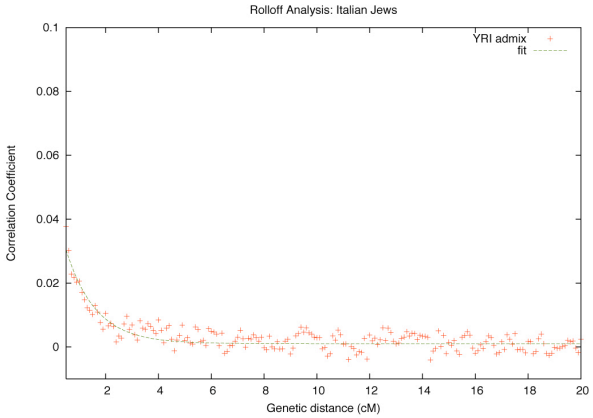








D. Jewish Groups



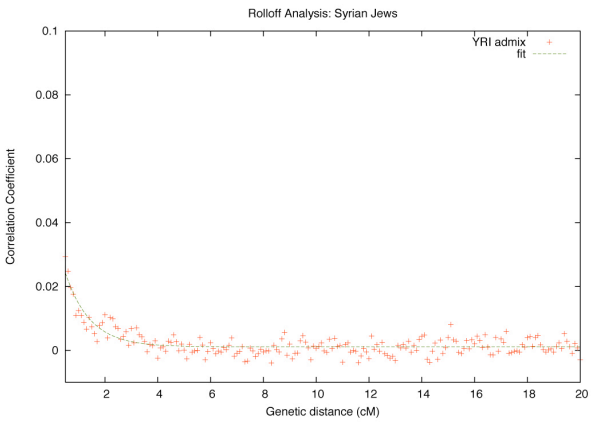


Figure A.10 (Continued)

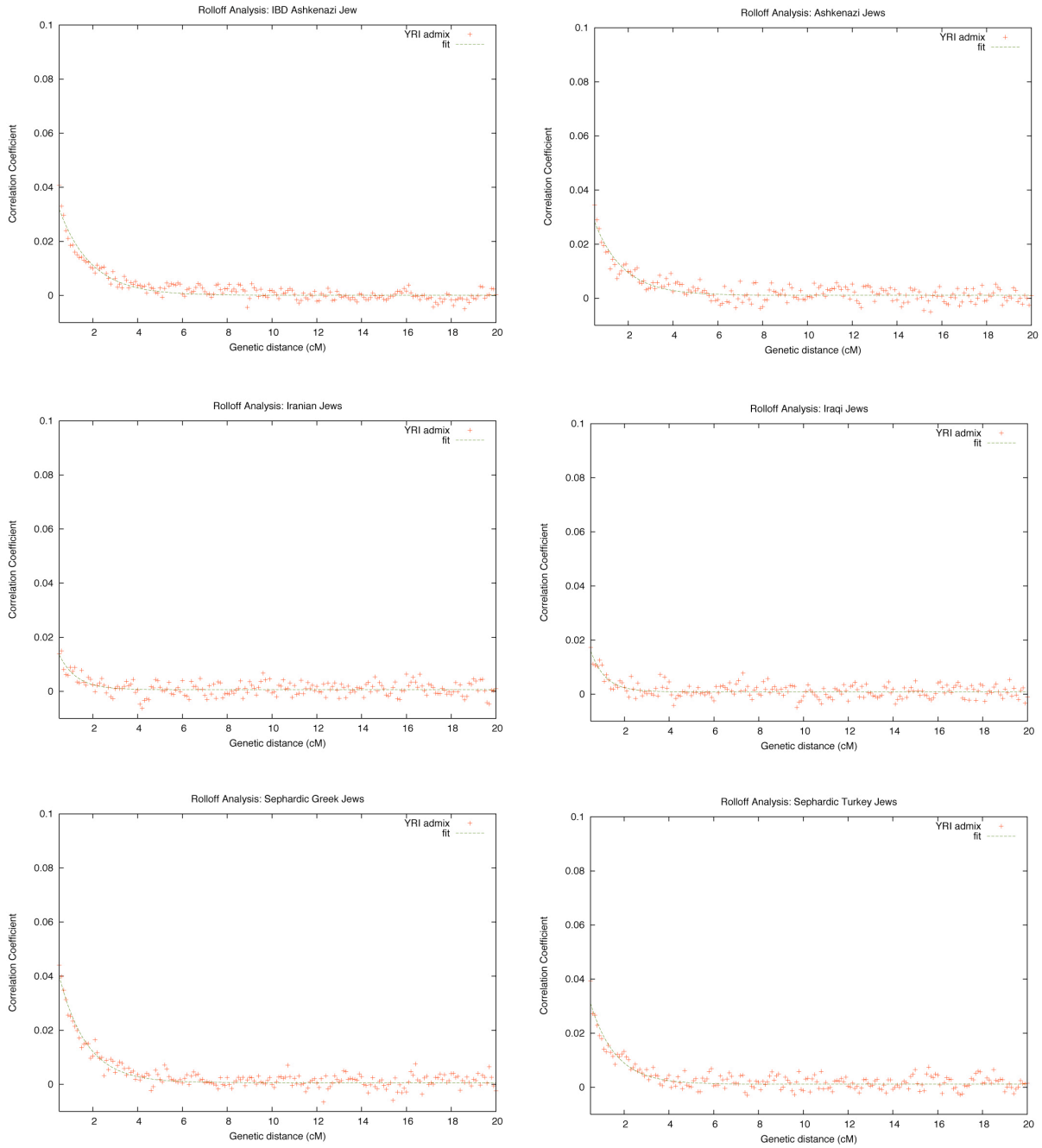


Figure A.10 (Continued) We performed *ROLLOFF* analysis for each West Eurasian population *X* that showed significant evidence of admixture in the *4 Population Test* using YRI and CEU as reference populations. We plot the decay of admixture LD as a function of genetic distance and estimate the date of admixture by fitting an exponential distribution to the data. Standard errors were calculated using a *Weighted Block Jackknife* as described in the Materials and Methods.

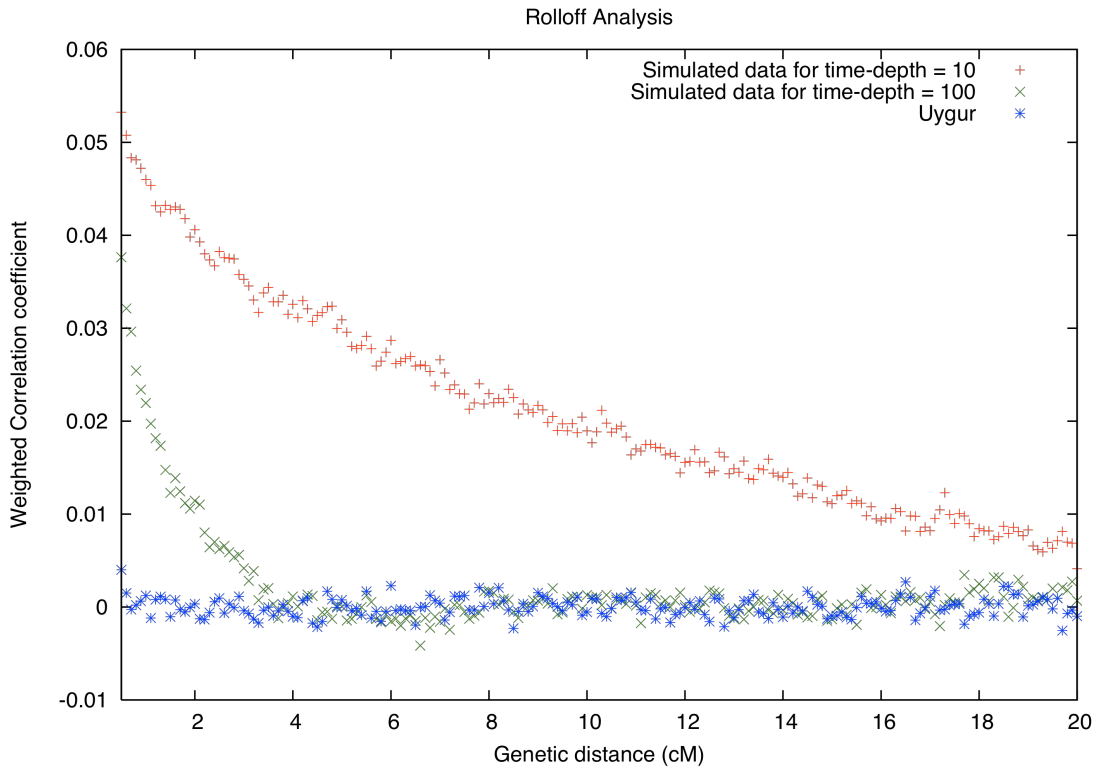


Figure A.11: *ROLLOFF* analysis in cases of no gene flow related to the tested ancestral populations. We performed *ROLLOFF* analysis for East Asian Uygurs, who have West Eurasians and Han Chinese ancestries. We used YRI and Pygmies (Mbuti and Biaka Pygmies) as the reference populations in *ROLLOFF* and saw no evidence of mixture. To show that this is not because of an inability to detect mixture when YRI and Pygmy-related groups are the true ancestral populations, we simulated 10 individuals of mixed Pygmy and Yoruba ancestry, with Yoruba mixture proportion (θ) = 80% and time since mixture (λ) = 10 generations (10 individuals) and θ = 80% and λ= 100 generations (10 individuals). We plot the *ROLLOFF* weighted correlation coefficient against genetic distance and observe clear evidence of mixture in these samples, with fairly accurately estimated dates of 10 and 90 generations respectively.

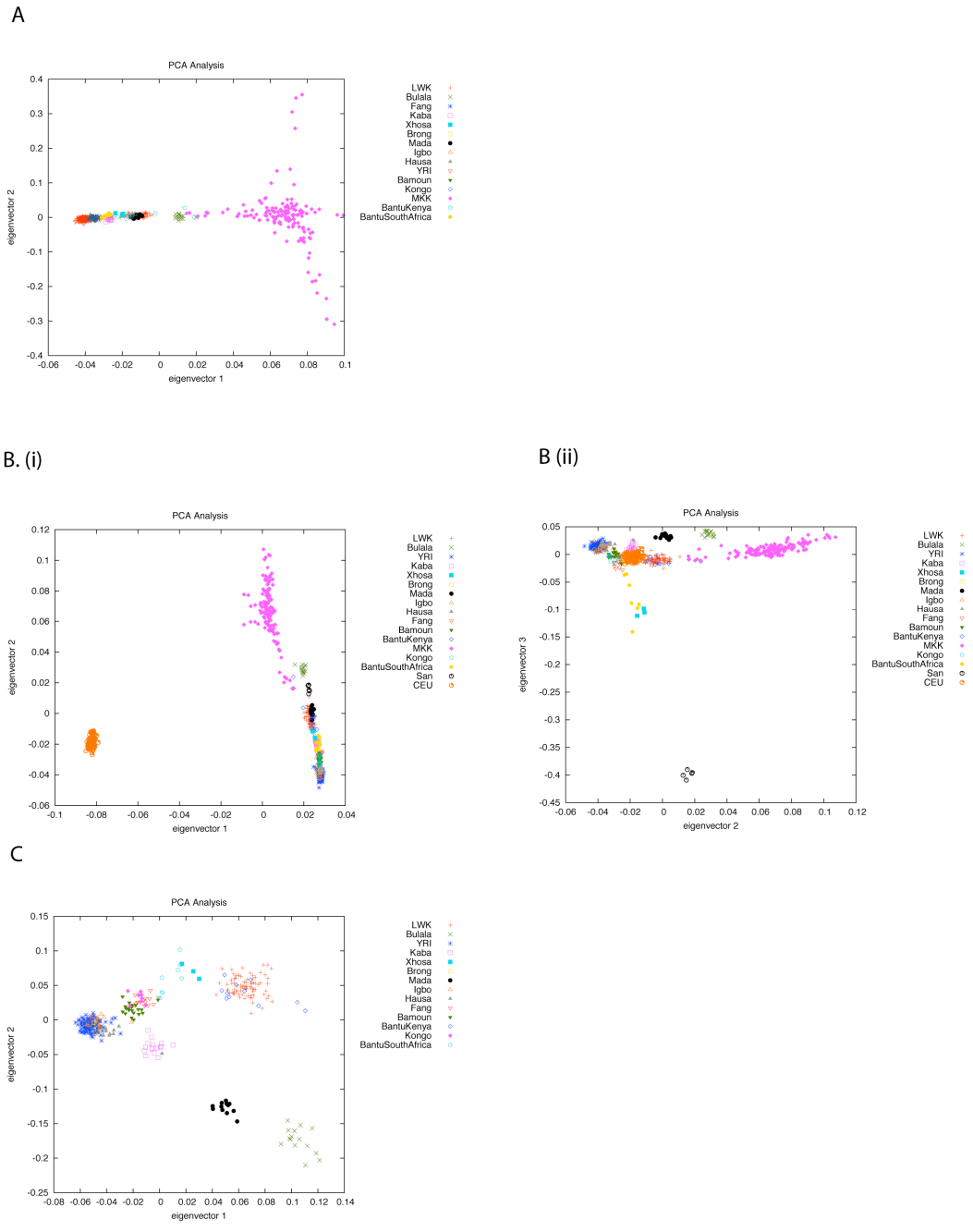


Figure A.12: Establishment of the axes of variation within Africa using PCA. To study the relationship of sub-Saharan African populations to each other and filter out populations with West Eurasian ancestry, we performed the following three PCA: (A) PCA of 15 sub-Saharan African groups using EIGENSOFT (B) PCA of 15 sub-Saharan African groups along with HapMap Chinese (CHB) and South African San, and (C) We PCA of 14 sub-Saharan African groups (excluding Kenyan Maasai).

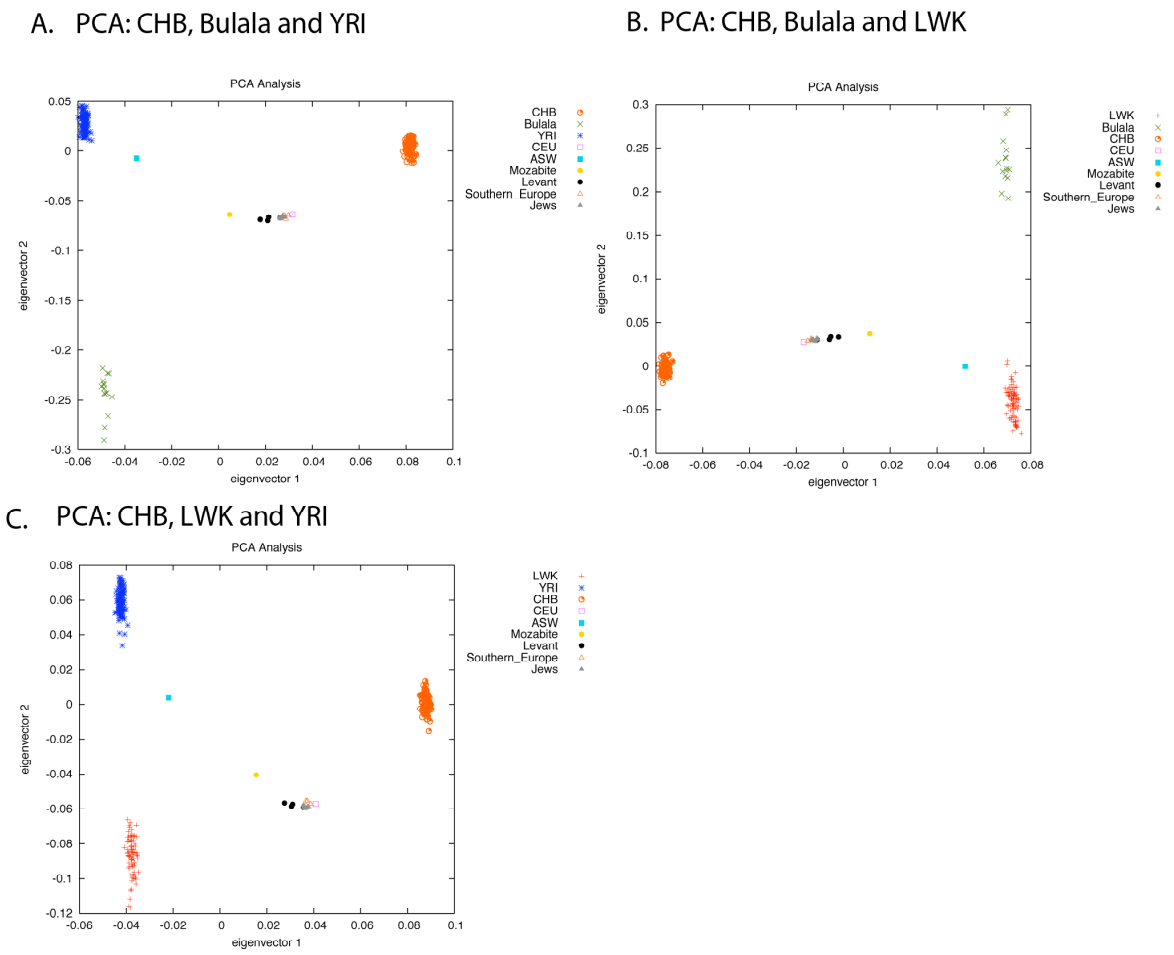


Figure A.13: Source of African ancestry in West Eurasians is likely to include some East African ancestors. In order to identify the source of the African ancestry in Levantines, Southern Europeans and Jews, we performed PCA Projection with all three possible pairs of African populations (Bulala, Kenyan Luhya (LWK) and Yoruba (YRI)) along with HapMap Chinese (CHB), and then plotted the mean values of all the samples from each West Eurasian population, African Americans (ASW) and North African (Mozabites) onto the first PC and second PC. These admixed West Eurasian populations align along a gradient that points more at (a) YRI than Bulala, and (b) LWK than Bulala suggesting little evidence for Chadic ancestry in West Eurasians. (c) Interesting, the PCA detects more relatedness to LWK than to YRI, suggesting that there may be some East African related ancestry in these West Eurasian populations.

Table A.1: Summary of Datasets

Merged Dataset	Dataset included	#Inds	#SNPs	Analyses in which this data set is used
A	POPRES	3,845	~500,000	
B	HGDP-CEPH	940	~650,000	
C	HapMap 3	1,115	>1M	
D	IBD	392	~300,000	
E	Jewish HapMap	232	~1M	
F	POPRES + HapMap3	4,960	347,315	*3 Population Test*
				ROLLOFF Analysis
				PCA Analysis
G	HGDP-CEPH + HapMap3	2,055	606,071	*3 Population Test*
				ROLLOFF Analysis
				PCA Analysis
H	IBD + HapMap3	1,507	284,951	*3 Population Test*
				ROLLOFF analysis
				PCA Analysis
I	Jewish HapMap +HapMap3	1,347	466,580	*3 Population Test*
				ROLLOFF analysis
				PCA Analysis
J	POPRES + HGDP-CEPH + HapMap3	5,900	85,628	*4 Population Test*
				f_4 Ancestry Estimation
K	IBD + HGDP-CEPH + HapMap3	2,447	284262	*4 Population Test*
				f_4 Ancestry Estimation
L	Jewish HapMap +HapMap3 + HGDP-CEPH	2,287	118,364	*f_4 Ancestry Estimation*
				4 Population Test
M	Jewish HapMap +HapMap3 + HGDP-CEPH + IBD + POPRES	3,614	36,175	*PCA Projection*

Table A.2: Outlier samples removed based on PCA curation

Population (X)	Dataset	Number of outliers removed
Bedouin	HGDP-CEPH	1
Swiss-Italian	POPRES	3
Swiss-French	POPRES	1
Palestine	HGDP-CEPH	3
Druze	HGDP-CEPH	1
Germany	POPRES	1
Croatia	POPRES	2
Greece	POPRES	8
Italy*	POPRES	24
Sardinia	POPRES	1
Tuscany	POPRES	1
Ashkenazi Jews*	IBD	69
Iraqi Jews	Jewish HapMap	1
Iranian Jews	Jewish HapMap	4
Italian Jews	Jewish HapMap	10
Sephardic Greek Jews	Jewish HapMap	7
Sephardic Turkey Jews	Jewish HapMap	3

* Due to evidence of population sub-structure, many individuals were excluded from these populations so that we were left with populations that were homogeneous in PCA.

Table A.3: Comparison of test statistics before and after PCA-based curation

Population (X)	Dataset	*Before data cleaning*			*After data cleaning*		
		Sam-ples (N)	West African ancestry proportion ± standard error	Estimated date of admixture (generations) ± standard error	Sam-ples (n)	West African ancestry proportion ± standard error	Estimated date of admixture (generations) ± standard error
African American	HapMap3	49	79.4% ± 0.3%	6 ± 1	49	79.4% ± 0.3%	6 ± 1
Palestinian	HGDP-CEPH	46	10.1% ± 0.4%	29 ± 2	43	9.3% ± 0.4%	34 ± 2
Bedouin	HGDP-CEPH	46	13.0% ± 0.4%	33 ± 4	--	--	--
Bedouin-g1*	HGDP-CEPH	--	--	--	15	14.5% ± 0.4%	34 ± 3
Bedouin-g2*	HGDP-CEPH	--	--	--	30	10.1% ± 0.4%	33 ± 3
Druze	HGDP-CEPH	41	4.2% ± 0.4%	53 ± 6	40	4.4% ± 0.4%	54 ± 7
Spain	POPRES	137	2.4% ± 0.3%	55 ± 3	137	2.4% ± 0.3%	55 ± 3
Portugal	POPRES	134	3.2% ± 0.3%	45 ± 5	134	3.2% ± 0.3%	45 ± 5
Italy	POPRES	225	2.2% ± 0.3%	68 ± 7	--	--	--
Swiss-Italian	POPRES	13	1.6% ± 0.5%	117 ± 41	--	--	--
Sardinian	HGDP-CEPH	28	2.9% ± 0.4%	95 ± 26	27	2.9% ± 0.4%	96 ± 28
Bergamo	HGDP-CEPH	12	1.6% ± 0.4%	214 ± 58	--	--	--
Tuscan	HGDP-CEPH	8	1.3% ± 0.4%	81 ± 62	--	--	--
Southern-Italy*	POPRES	--	--	--	121	2.7% ± 0.3%	62 ± 6
Northern-Italy*	POPRES	--	--	--	90	1.1% ± 0.3%	154 ± 27
Swiss-French	POPRES	760	0.5% ± 0.2%	71 ± 5	759	0.5% ± 0.2%	71 ± 6
Ashkenazi Jews	IBD	392	2.8% ± 0.3%	54 ± 8	323	2.8% ± 0.3%	91 ± 11
Ashkenazi Jews	Jewish HapMap	34	3.2% ± 0.4%	76 ± 13	34	3.2% ± 0.4%	76 ± 13
Syrian Jews	Jewish HapMap	25	3.9% ± 0.5%	99 ± 23	25	3.9% ± 0.5%	99 ± 23
Iranian Jews	Jewish HapMap	28	2.4% ± 0.6%	127 ± 44	24	2.6% ± 0.6%	129 ± 34
Iraqi Jews	Jewish HapMap	37	3.8% ± 0.5%	155 ± 20	36	3.8% ± 0.5%	153 ± 22
Sephardic Greek Jews	Jewish HapMap	39	4.8% ± 0.4%	82 ± 8	39	4.8% ± 0.4%	82 ± 8
Sephardic Turkey Jews	Jewish HapMap	32	4.1% ± 0.4%	84 ± 11	27	4.5% ± 0.4%	89 ± 11
Italian Jews	Jewish HapMap	37	4.7% ± 0.5%	92 ± 18	27	4.9% ± 0.5%	88 ± 19

Note: * indicates that new population label based on PCA curation. Mixture proportion estimates are based on f_4 *Ancestry Estimation* using San, Yoruba, CEU and Papuan as the reference populations. The *ROLLOFF* estimated date of mixture uses CEU and YRI as the proposed ancestral populations

Table A.4: *4 Pop Test* using different ancestral pops. compared to Table 2.1

			Z*-score for *4 Population test		
Population (X)	**Region**	**Dataset**	**$(P_{Papuan}-P_{YRI}) \times (P_{Adygei}-P_X)$**	**$(P_{Papuan}-P_{Mandenka}) \times (P_{CEU}-P_X)$**	**$(P_{Papuan}-P_{BantuKenya}) \times (P_{CEU}-P_X)$**
African American	n/a	HapMap3	**-85.6**	**-85.1**	**-77.7**
Palestine	L	HGDP-CEPH	**-28.1**	**-27.3**	**-27.8**
Turkey	L	POPRES	-2.2	-1.4	-1.4
Bedouin-g1	L	HGDP-CEPH	**-37.3**	**-35.6**	**-35.4**
Bedouin-g2	L	HGDP-CEPH	**-25.7**	**-25.7**	**-26**
Druze	L	HGDP-CEPH	**-16.1**	**-14.5**	**-14.8**
Spain	SE	POPRES	**-8**	**-12.4**	**-11.7**
Portugal	SE	POPRES	**-10**	**-14.6**	**-14.4**
Romania	SE	POPRES	-2	-1	-0.9
Croatia	SE	POPRES	-0.6	0.7	0.9
Bosnia-Herzegovina	SE	POPRES	-1.8	-0.5	-0.5
Sardinia	SE	HGDP-CEPH	**-9.3**	**-9.5**	**-10**
Southern-Italy	SE	POPRES	**-9.2**	**-11.2**	**-10.9**
Northern-Italy	SE	POPRES	**-5**	**-6**	**-5.8**
Austria	ECE	POPRES	-1.7	-0.5	-0.4
Poland	ECE	POPRES	-0.7	1.4	1.2
Hungary	ECE	POPRES	-1.5	0.1	0.4
Czech Republic	ECE	POPRES	-1.1	-0.3	0.6
Adygei	ECE	HGDP-CEPH	--	3.1	2.7
Russia	ECE	POPRES	-0.7	0.7	0.5
Russia	ECE	HGDP-CEPH	-0.7	0.7	0.5
Swiss-French	I	POPRES	**-3.3**	**-3.6**	**-3.5**
France	I	POPRES	-2.8	-2.2	-2.4
France	I	HGDP-CEPH	-2.8	-2.2	-2.4
Basque	I	HGDP-CEPH	-3	-1.5	-1.5
Belgium	I	POPRES	-2.3	-1	-1
Orkney	I	POPRES	-0.2	**3.2**	2.9
United Kingdom	I	POPRES	-1.4	1.1	1.1
Ireland	I	POPRES	-0.9	2	1.9
Scotland	I	POPRES	1.6	3	2.9
Netherlands	I	POPRES	-0.9	1	0.8
Swiss-German	I	POPRES	-2.5	-1.5	-1.3
Germany	I	POPRES	-2.3	-1	-0.8
Sweden	I	POPRES	-0.2	1.9	1.7
Ashkenazi Jews	n/a	IBD	**-11.4**	**-11.7**	**-11.4**
Ashkenazi Jews	n/a	Jewish HapMap	**-8.8**	**-9.6**	**-9.5**
Syrian Jews	n/a	Jewish HapMap	**-10.1**	**-10.2**	**-10**
Iranian Jews	n/a	Jewish HapMap	**-6.6**	**-5.9**	**-5.7**
Iraqi Jews	n/a	Jewish HapMap	**-9.4**	**-8.8**	**-8.8**
Sephardic Greek Jews	n/a	Jewish HapMap	**-12.4**	**-13.8**	**-13.8**
Sephardic Turkey Jews	n/a	Jewish HapMap	**-11.9**	**-13.6**	**-13.5**
Italian Jews	n/a	Jewish HapMap	**-10.8**	**-11.3**	**-11.6**

Note: We analyzed data from all West Eurasian populations with at least 5 samples. Regions are abbreviated as: I – Northwest Europe, ECE – East-Central Europe, SE – Southern Europe and L – Levant. For the *4 Population Test,* we report only results for the tree shown in the table. Results for all alternate topologies show even higher violations of the tree ($|Z| >> 15$). Scores that are significant are highlighted in bold.

Table A.5: *f_4 Ancestry Estimation* using different ancestral populations compared to Table 2.2

Population (X)	Dataset	Region	(San, (Mandenka, (Papuan, (CEU,X))))	(San,(BantuKenya, (Papuan,(CEU,X))))	(San,(YRI, (CHB, (CEU,X))))
African Americans	HapMap3	n/a	81.0% ± 0.3%	82.3% ± 0.4%	79.0% ± 0.2%
Palestinian	HGDP-CEPH	L	9.5% ± 0.4%	9.6% ± 0.4%	10.3% ± 0.3%
Bedouin-g1	HGDP-CEPH	L	14.8% ± 0.4%	15.1% ± 0.4%	15.1% ± 0.3%
Bedouin-g2	HGDP-CEPH	L	10.3% ± 0.5%	10.4% ± 0.5%	11.0% ± 0.4%
Druze	HGDP-CEPH	SE	4.5% ± 0.4%	4.5% ± 0.4%	5.2% ± 0.3%
Spain	POPRES	SE	2.5% ± 0.3%	3.3% ± 0.3%	3.5% ± 0.2%
Portugal	POPRES	SE	3.2% ± 0.3%	2.5% ± 0.3%	2.8% ± 0.2%
Sardinian	HGDP-CEPH	SE	3.0% ± 0.4%	3.0% ± 0.4%	3.4% ± 0.3%
Southern-Italy	POPRES	SE	2.8% ± 0.3%	2.8% ± 0.3%	3.2% ± 0.3%
Northern-Italy	POPRES	SE	1.2% ± 0.3%	1.2% ± 0.3%	1.3% ± 0.3%
Swiss-French	POPRES	I	0.5% ± 0.2%	0.5% ± 0.2%	0.5% ± 0.2%
Ashkenazi Jews	IBD	n/a	3.3% ± 0.3%	2.9% ± 0.3%	2.9% ± 0.3%
Ashkenazi Jews	Jewish HapMap	n/a	3.2% ± 0.4%	2.9% ± 0.5%	3.0% ± 0.4%
Syrian	Jewish HapMap	n/a	3.9% ± 0.5%	4.0% ± 0.5%	4.6% ± 0.4%
Iranian	Jewish HapMap	n/a	2.6% ± 0.6%	2.7% ± 0.7%	2.9% ± 0.5%
Iraqi	Jewish HapMap	n/a	3.8% ± 0.6%	3.9% ± 0.6%	4.5% ± 0.4%
Sephardic Greek Jews	Jewish HapMap	n/a	4.9% ± 0.4%	4.9% ± 0.4%	5.3% ± 0.4%
Sephardic Turkey Jews	Jewish HapMap	n/a	4.6% ± 0.4%	4.6% ± 0.4%	5.5% ± 0.4%
Italian Jews	Jewish HapMap	n/a	5.0% ± 0.5%	5.0% ± 0.5%	5.2% ± 0.4%

Note: Estimates of proportions of mixture for all West Eurasians that give statistically signal evidence of mixture in Table 2.1. Regions are abbreviated as: I – Northwest Europe, SE – Southern Europe and L – Levant. Mixture proportions are based on *f_4 Ancestry Estimation* method using the phylogenetic trees specified in columns 4, 5 and 6.

Table A.6: Simulation to test the effect of ascertainment bias on *3 Pop. Test* Results

Model	Divergence time (t_{AB})	Effective pop. size of Pop *A* (N_A) and Pop *B* (N_B)	*3 Pop. Test* $(P_C-P_A)(P_C-P_B)$
1: One chromosome from each Pop *A* and Pop *B*	45,000	$N_A = N_o$, $N_B = 0.25N_o$	-48.8
	60,000	$N_A = N_o$, $N_B = 0.4N_o$	-54.7
	100,000	$N_A = N_o$, $N_B = 0.85N_o$	-37.3
2: Both chromosomes from Pop *A*	45,000	$N_A = N_o$, $N_B = 0.25N_o$	-68.6
	60,000	$N_A = N_o$, $N_B = 0.4N_o$	-90.8
	100,000	$N_A = N_o$, $N_B = 0.85N_o$	-45.1
3: Both chromosomes from Pop *B*	45,000	$N_A = N_o$, $N_B = 0.25N_o$	-40.8
	60,000	$N_A = N_o$, $N_B = 0.4N_o$	-20.4
	100,000	$N_A = N_o$, $N_B = 0.85N_o$	-40.6

NOTE: Details of the demographic model used for the simulation are shown in Figure A.4. Effective population size of Pop *B* (N_B) is set such that the $F_{ST}(A,B) = 0.15$

Table A.7: *ROLLOFF* Simulations: Effect of variations in bin sizes and genetic map

Category		Simulation at time depth (λ) = 10 generations	Simulation at time depth (λ) = 100 generations
Genetic Map	0.00001	11 ± 1	113 ± 5
Rate parameter (Φ)	0.0001	11 ± 1	113 ± 5
	0.001	11 ± 1	114 ± 5
	0.1	11 ± 1	113 ± 5
	10	11 ± 1	113 ± 5
Bin Size (cM)	0.01	11 ± 1	114 ± 6
	0.025	11 ± 1	114 ± 5
	0.1	11 ± 1	113 ± 5
	0.4	11 ± 1	115 ± 6
	1	13 ± 1	145 ± 4

Note: We simulated 10 individuals using YRI and CEU as the ancestral populations where we set the mixture proportion to be θ = 20% and the time since mixture to be λ= 10 or 100 generations. We then performed *ROLLOFF* analysis with an independent dataset of European Americans and Nigerian Yorubans as reference population. To test the effect of inaccuracies in the genetic map, we systematically change the genetic map by modeling the change based on the convolution property of a gamma distribution with rate parameter φ. A low value of φ implies significant changes to the map and a high value allows for fine scale changes. To test the effect of the bin size, we vary the bin size within the range of 0.01 - 1cM and test the effect on the estimated dates.

Table A.8*: ROLLOFF* simulations: Effect of inaccurate ancestral populations

Reference Populations	$F_{ST}$with ancestral pops. CEU and YRI	Estimated date for time depth λ =10 generations	Estimated date for time depth λ = 100 generations
European American	0.00	11 ± 1	113 ± 5
Yoruba	0.00		
Basque	0.01	11 ± 1	125 ± 8
Mandenka	0.01		
Druze	0.02	11 ± 1	124 ± 12
Yoruba	0.00		
Druze	0.02	11 ± 1	119 ± 12
Kenyan Bantu	0.01		
Gujarati	0.03	11 ± 1	112 ± 6
Maasai	0.03		

Note: We simulated 10 individuals using YRI and CEU as the ancestral populations where we set the mixture proportion to be θ = 20% and the time since mixture to be λ= 10 or 100 generations. We then performed *ROLLOFF* analysis with the reference populations shown in column 1. Average allele frequency difference (F_{ST}) between the true ancestral population (CEU and YRI) and the reference population used for *ROLLOFF* analysis is shown in column 2

Table A.9: *ROLLOFF* simulations: Effect of inaccurate ancestral populations in the case of low mixture proportions and old mixture dates.

Reference Populations used in *ROLLOFF*	F_{ST} with true ancestral populations	Simulated date given a truth of l=55 generations (bias in parentheses)			Simulated date given a truth of l=89 generations (bias in parentheses)		
		θ= 1%	θ= 2%	θ= 3%	θ= 3%	θ= 4%	θ= 5%
European American (*n*=100)	0.00	68 (24%)	62 (13%)	59 (8%)	98 (10%)	98 (10%)	95 (6%)
Yoruba (*n*=100)	0.00						
Basque (*n*=24)	0.01	71 (29%)	63 (15%)	60 (10%)	101 (13%)	100 (15%)	95 (7%)
Mandenka (*n*=22)	0.01						
Druze (*n*=42)	0.02	74 (35%)	63 (14%)	61 (10%)	102 (15%)	100 (12%)	96 (8%)
Yoruba (*n*=21)	0.00						
Druze (*n*=42)	0.02	75 (36%)	63 (14%)	61 (11%)	103 (16%)	100 (13%)	97 (9%)
Kenyan Bantu (*n*=11)	0.01						
Gujarati (*n*=88)	0.03	75 (36%)	66 (20%)	61 (12%)	103 (16%)	102 (15%)	97 (9%)
Maasai (*n*=143)	0.03						

Note: We simulated 20 individuals with CEU and YRI as the ancestral populations where we set the mixture proportion (θ), time since mixture (λ) and the reference populations used for the *ROLLOFF* analysis as shown above. We repeated each simulation 100 times and estimated the average and bias for each run. The values shown in the rows are the average for the 100 simulations and bias, defined as (average-truth)/(truth).

Table A.10: *ROLLOFF* simulations: Effect of number of admixed samples

# of admixed samples	Average estimated date (bias in simulations)
n=10	66 (22%)
n=20	56 (4%)
n=30	56 (4%)
n=40	57 (6%)
n=50	55 (2%)
n=80	54 (0%)
n=100	54 (0%)

Note: We simulated *n* individuals using European Americans and Nigerians as the ancestral populations where we set the mixture proportion to be θ=2% and the time since mixture to be λ= 54 generations, and then performed *ROLLOFF* analysis with HapMap3 CEU and YRI as the reference populations. We repeated each simulation 100 times and estimated the average and bias. The values shown in the cells are the average over the 100 simulations and the bias, defined as (average-truth)/(truth).

Table A.11: *ROLLOFF* simulations: Effect of mixture proportions

Mixture proportion	Average Estimated date (bias in simulations)
θ=1%	62 (15%)
θ=2%	58 (7%)
θ=3%	57 (6%)
θ=5%	55 (2%)
θ=10%	54 (0%)
θ=20%	53 (2%)
θ=30%	53 (2%)
θ=50%	53 (2%)

Note: We simulated 50 individuals using YRI and CEU as the ancestral populations where we set the mixture proportion to be θ (shown in the table) and the time since mixture to be λ= 54 generations. We then performed *ROLLOFF* analysis with an independent dataset of 100 European Americans and 100 Nigerian Yorubans as reference population. We repeated each simulation 100 times and estimated the average and bias. The bias is defined as (average-truth)/(truth).

Table A.12: *ROLLOFF* analysis of West Eurasians: bias in the estimated date for empirically estimated parameters

Population (X)	Dataset	Samples	West African ancestry proportion ± standard error	Estimated date of admixture (generations) +/- standard error	Simulation Results Average (bias in simulations)
African American	HapMap3	49	79.4% ± 0.3%	6 ± 1	6 (0%)
Palestinian	HGDP-CEPH	43	9.3% ± 0.4%	34 ± 2	35 (3%)
Bedouin-g1	HGDP-CEPH	15	14.5% ± 0.4%	34 ± 3	36 (6%)
Bedouin-g2	HGDP-CEPH	30	10.1% ± 0.4%	33 ± 2	35 (6%)
Druze	HGDP-CEPH	41	4.4% ± 0.4%	54 ± 7	64 (19%)
Spain	POPRES	137	2.4% ± 0.3%	55 ± 3	55 (0%)
Portugal	POPRES	134	3.2% ± 0.3%	45 ± 5	45 (0%)
Sardinian	HGDP-CEPH	27	2.9% ± 0.5%	96 ± 28	121 (26%)
Southern-Italy	POPRES	121	2.7% ± 0.3%	62 ± 6	62 (0%)
Northern-Italy	POPRES	90	1.1% ± 0.3%	154 ± 27	128 (-17%)
Swiss-French	POPRES	759	0.5% ± 0.2%	71 ± 6	n/a
Ashkenazi Jews	IBD	323	2.8% ± 0.3%	91 ± 11	n/a
Ashkenazi Jews	Jewish HapMap	34	3.2% ± 0.4%	76 ± 13	99 (31%)
Syrian Jews	Jewish HapMap	25	3.9% ± 0.5%	99 ± 23	126 (27%)
Iranian Jews	Jewish HapMap	24	2.6% ± 0.6%	129 ± 34	188 (46%)
Iraqi Jews	Jewish HapMap	36	3.8% ± 0.5%	153 ± 22	191 (25%)
Sephardic Greek Jews	Jewish HapMap	39	4.8% ± 0.4%	82 ± 8	102 (24%)
Sephardic Turkey Jews	Jewish HapMap	27	4.5% ± 0.4%	89 ± 11	105 (18%)
Italian Jews	Jewish HapMap	27	4.9% ± 0.5%	88 ± 19	103 (17%)

Note: We simulated individuals of mixed European and African ancestry where we set the sample size, mixture proportion (θ) and time since mixture (λ) to match the parameters estimated for West Eurasians. We then performed *ROLLOFF* analysis using HapMap3 Italian Toscanis (TSI) and Kenyan Luhya (LWK) as reference populations (as data for the true ancestral populations might not always be available for real samples). We repeated each simulation 100 times and estimated the average and bias. Average = mean of the estimated date for 100 simulations and Bias = (average-truth)/(truth). We were not able to perform simulations for Swiss-French and IBD Ashkenazi Jews because we did not have a sufficiently large pool of ancestral haplotypes to accommodate the large sample sizes in these groups.

Table A.13: *ROLLOFF* Simulations: Continuous admixture scenarios

	Δ =1	**Δ =5**	**Δ =10**	**Δ =20**	**Δ =30**	**Δ =100**
***a* =1**	1 ± 0	4 ± 0	4 ± 1	6 ± 1	11 ± 2	30 ± 6
***a* =25**	26 ± 2	25 ± 2	30 ± 3	33 ± 2	41 ± 2	63 ± 6
***a* =50**	55 ± 3	56 ± 3	52 ± 3	59 ± 5	60 ± 5	77 ± 6
***a* =100**	100 ± 6	117 ± 7	109 ± 7	102 ± 8	118 ± 5	142 ± 13
***a* =200**	204 ± 18	188 ± 19	202 ± 17	253 ± 27	198 ± 19	268 ± 25

Note: We performed 30 simulations with CEU and YRI as the ancestral populations using the simulation method described in Note S3f. In each simulation, we varied the length of the interval *I (Δ = b-a)* during which YRI lineages migrate to the CEU population (creating an admixed population). Following the YRI mixture, there are '*a*' generations of random mixture between the admixed individuals. This can roughly be thought of as simulating genetic drift, since admixture. We performed *ROLLOFF* analysis using a non-overlapping dataset of 1,107 European American and 737 Nigerian Yoruba individuals as reference samples.

Table A.14: *4 Population Test* to distinguish between East & West African ancestry

Pop X	Dataset	Region	$(P_{CEU}-P_X)$ $(P_{YRI}-P_{LWK})$
African Americans	HapMap3	n/a	**-46.8**
Mozabite	HGDP-CEPH	n/a	**-24.5**
Palestinian	HGDP-CEPH	L	**-5.1**
Bedouin-g1	HGDP-CEPH	L	**-7.2**
Bedouin-g2	HGDP-CEPH	L	-1.9
Druze	HGDP-CEPH	L	0.4
Spain	POPRES	SE	-2.5
Portugal	POPRES	SE	-2.9
Northern Italy	POPRES	SE	0.9
Southern Italy	POPRES	SE	1.3
Sardinian	HGDP-CEPH	SE	1.9
Swiss-French	POPRES	I	1.0
Ashkenazi Jews	IBD	n/a	0.0
Ashkenazi Jews	Jewish HapMap	n/a	0.8
Iranian Jews	Jewish HapMap	n/a	1.0
Iraqi Jews	Jewish HapMap	n/a	2.1
Italian Jews	Jewish HapMap	n/a	1.1
Sephardic Greek Jews	Jewish HapMap	n/a	1.0
Sephardic Turkey Jews	Jewish HapMap	n/a	0.6
Syrian Jews	Jewish HapMap	n/a	0.9

Note: We analyzed data from all West Eurasian populations that showed evidence of African ancestry in Table 2.1. Regions are abbreviated as: I – Northwest Europe, SE – Southern Europe and L – Levant. For the *4 Population Test,* we report only results for the tree shown in the table. Results for all alternate topologies show even higher violations of the tree ($|Z| >> 15$). Scores that are significant are highlighted in bold.

Table A.15: Estimated mixture proportion and date using East Africans as reference population

Population (X)	Dataset	Region	Samples	East African ancestry proportion ± standard error	Estimated date of admixture (generations) +/- standard error
African Americans	HapMap3	n/a	49	81.1% ± 0.3%	6 ± 1
Palestinian	HGDP-CEPH	L	43	10.6% ± 0.3%	33 ± 2
Bedouin-g1	HGDP-CEPH	L	15	15.5% ± 0.4%	34 ± 3
Bedouin-g2	HGDP-CEPH	L	30	11.3% ± 0.4%	34 ± 3
Druze	HGDP-CEPH	L	41	5.4% ± 0.3%	56 ± 7
Spain	POPRES	SE	137	2.9% ± 0.2%	57 ± 3
Portugal	POPRES	SE	134	3.6% ± 0.2%	46 ± 5
Sardinian	HGDP-CEPH	SE	27	3.5% ± 0.3%	93 ± 30
Southern-Italy	POPRES	SE	121	3.3% ± 0.3%	62 ± 6
Northern-Italy	POPRES	SE	90	1.3% ± 0.3%	158 ± 31
Swiss-French	POPRES	I	759	0.5% ± 0.2%	74 ± 6
Ashkenazi Jews	IBD	n/a	323	3.0% ± 0.3%	89 ± 12
Ashkenazi Jews	Jewish HapMap	n/a	34	3.3% ± 0.5%	77 ± 14
Syrian Jews	Jewish HapMap	n/a	25	4.0% ± 0.5%	103 ± 23
Iranian Jews	Jewish HapMap	n/a	24	2.7% ± 0.7%	114 ± 30
Iraqi Jews	Jewish HapMap	n/a	36	3.9% ± 0.6%	149 ± 23
Sephardic Greek Jews	Jewish HapMap	n/a	39	4.9% ± 0.4%	80 ± 7
Sephardic Turkey Jews	Jewish HapMap	n/a	27	4.6% ± 0.4%	91 ± 11
Italian Jews	Jewish HapMap	n/a	27	5.0% ± 0.5%	89 ± 18

Note: Regions are abbreviated as: I – Northwest Europe, SE – Southern Europe and L – Levant. Mixture proportion estimates are based on f_4 *Ancestry Estimation* using San, LWK, CEU and Papuan as the reference populations. The *ROLLOFF* estimated date of mixture uses CEU and LWK as the reference populations.

Table A.16: *ROLLOFF* Analysis for different jackknife block sizes: example Spain

Jackknife Block size	Estimated date ± standard error
1 chromosome	55 ± 3
5cM	55 ± 3
10cM	55 ± 3
20cM	55 ± 3

NOTE: The *ROLLOFF* estimated date of mixture uses CEU and YRI as the reference populations.

References

1. Sun, J., Mullikin, J., Patterson, N., and Reich, D. (2009). Microsatellites are molecular clocks that support accurate inferences about history. Molecular biology and evolution 26, 1017.

2. Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2, e190.

3. Reich, D., Thangaraj, K., Patterson, N., Price, A., and Singh, L. (2009). Reconstructing Indian population history. Nature 461, 489-494.

4. Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337.

5. Smith, M., Patterson, N., Lautenberger, J., Truelove, A., McDonald, G., Waliszewska, A., Kessing, B., Malasky, M., Scafe, C., and Le, E. (2004). A high-density admixture map for disease gene discovery in african americans. The American Journal of Human Genetics 74, 1001-1013.

6. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. In. (American Association for the Advancement of Science), pp 321-324.

7. Price, A., Tandon, A., Patterson, N., Barnes, K., Rafaels, N., Ruczinski, I., Beaty, T., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genetics 5.

8. Xu, S., and Jin, L. (2008). A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. The American Journal of Human Genetics 83, 322-336.

9. Li, J., Absher, D., Tang, H., Southwick, A., Casto, A., Ramachandran, S., Cann, H., Barsh, G., Feldman, M., and Cavalli-Sforza, L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100.

10. Bryc, K., Auton, A., Nelson, M., Oksenberg, J., Hauser, S., Williams, S., Froment, A., Bodo, J., Wambebe, C., and Tishkoff, S. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences 107, 786.

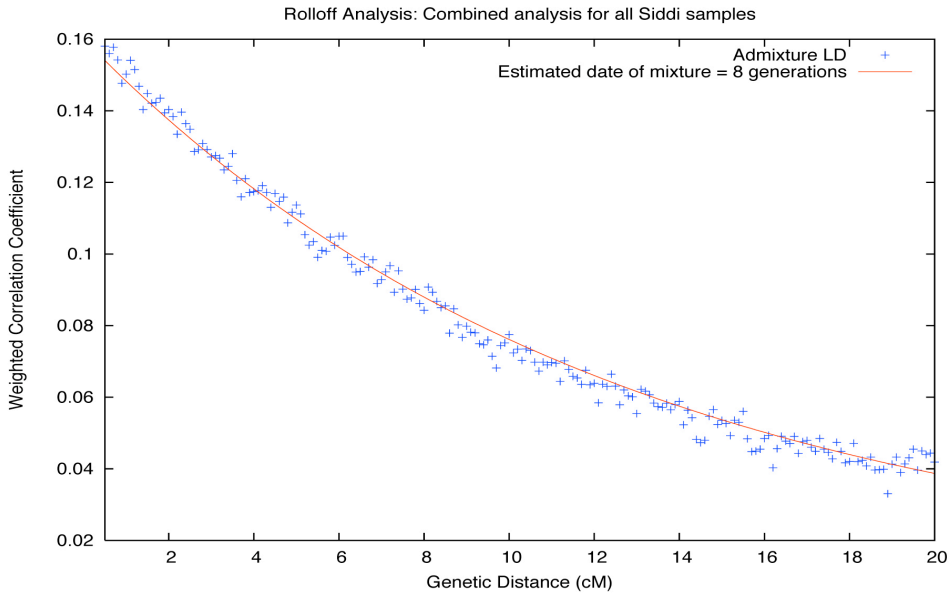
11. Altshuler, D., Brooks, L., Chakravarti, A., Collins, F., Daly, M., and Donnelly, P. (2005). A haplotype map of the human genome. Nature 437, 1299-1320.

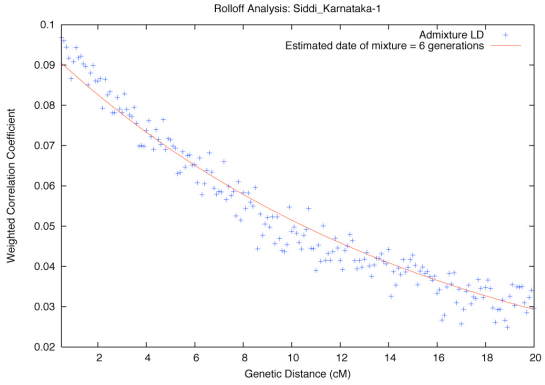
12. Gibbon, E. (1890). The decline and fall of the Roman Empire.(WW Gibbings).

13. Richards, M., Rengo, C., Cruciani, F., Gratrix, F., Wilson, J., Scozzari, R., Macaulay, V., and Torroni, A. (2003). Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. The American Journal of Human Genetics 72, 1058-1064.

Appendix B

Supplementary Material for Chapter 3





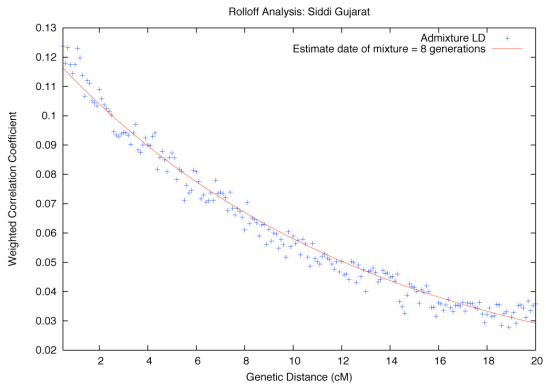


Figure B.1: *ROLLOFF* Analysis using East Africans as the ancestral population. We performed the *ROLLOFF* analysis to estimate the date of admixture in Siddis, using East Africans (LWK) and ICP samples as the reference populations. The estimated dates of mixture are as following: Siddi-Karnataka-1 = 6 ± 1 generation, Siddi-Gujarat = 8 ± 1 generation and the result for the combined data set (n = 12) is 8 ± 1 generation or ~200 years. Due to limited number of samples, we were not able to perform separate analysis of the Siddi_Karnataka-2 group.

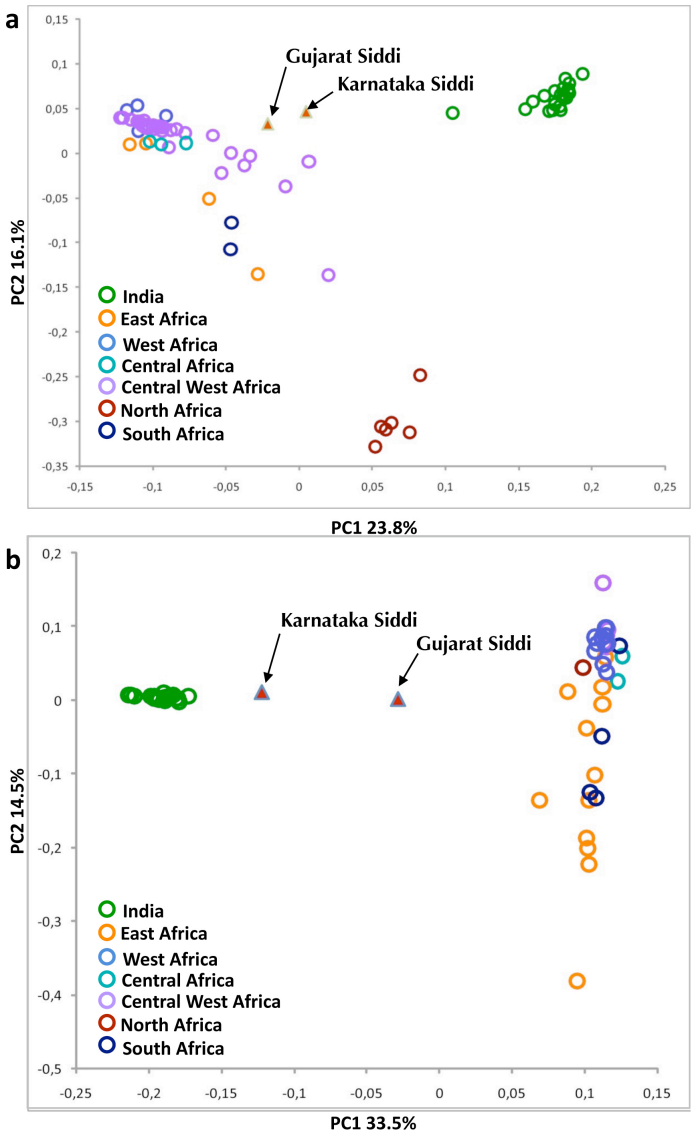


Figure B.2: Principal Component Analysis (PCA) using data for uniparentally transmitted markers. PCA was performed using Y-chromosome and mitochondrial DNA haplogroups frequencies among Siddis, Indian and African populations. (a) PCA based on Y-chromosome haplogroup frequencies (Supporting Dataset S1) shows that the Siddis from Gujarat and Karnataka are related to African populations and nearby Indian populations. The Siddis also appear to be most closely related to the Bantu speaking populations from West and Central West Africa. (b) PCA based on mitochondrial DNA (mtDNA) haplogroup frequencies (Supporting Dataset S2) shows the Siddi falling between the two major clusters on PC1 of Indians and Africans. The Siddis from Karnataka state appear to be closer to the Indians than Africans, likely because of high level of admixture with the nearby Indian groups.

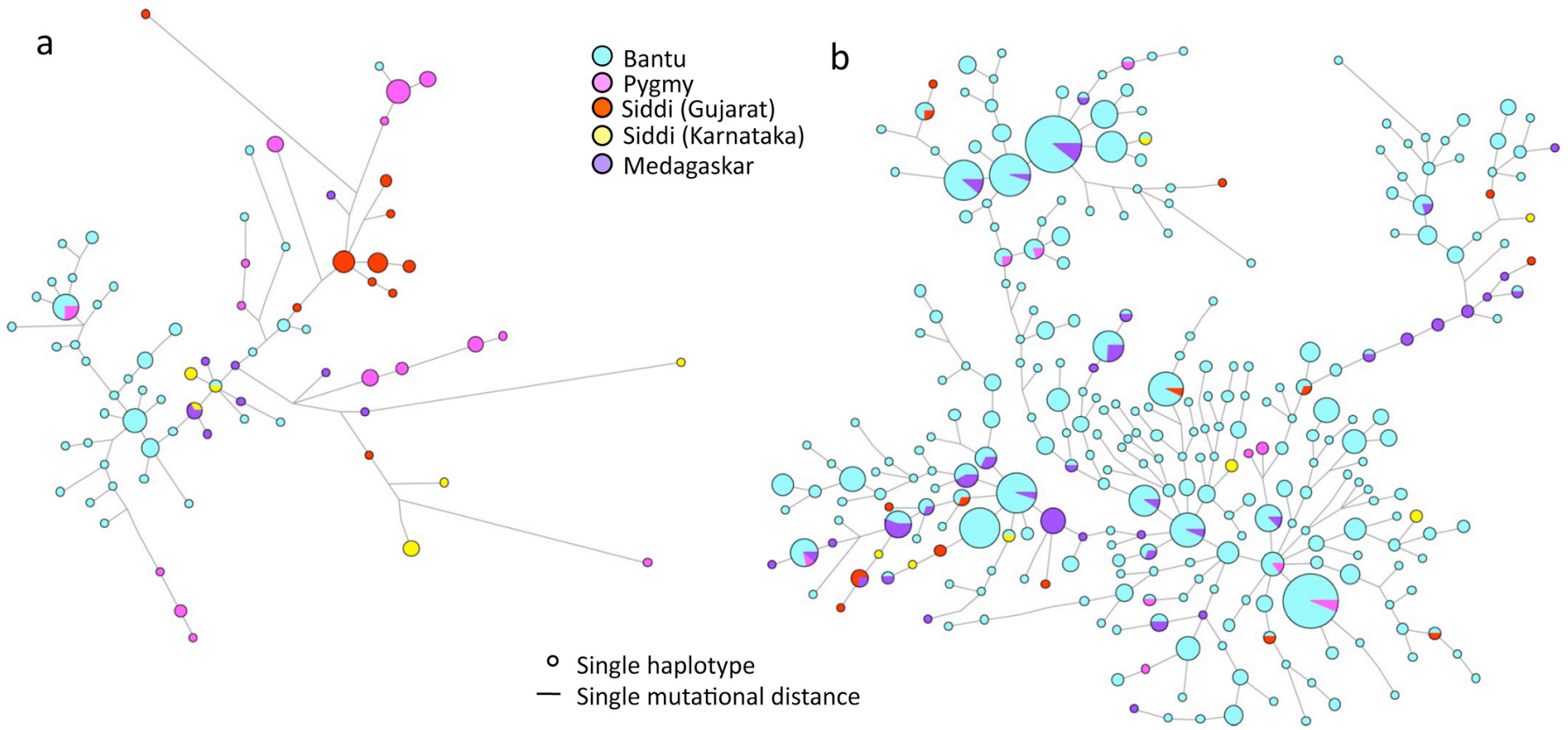


Figure B.3: Phylogenetic Network analysis. Phylogenetic networks were constructed based on ten common Y-chromosomal short tandem repeats (Y-STRs). (a) Haplotype networks were generated for haplogroups B2 and (b) E1b1 by using the median-joining algorithm of Network 4.6 (http://www.fluxus-engineering.com). Because of the high level of reticulation in the E1b1a-M2 samples, data were post-processed using the MP calculation option in Network 4.6. The size of the circles is proportional to the number of samples. Published data from Berniell-Lee et al. 2009 and Tofanelli et al. 2009 was used for Bantu, Pygmy and Madagascar populations.

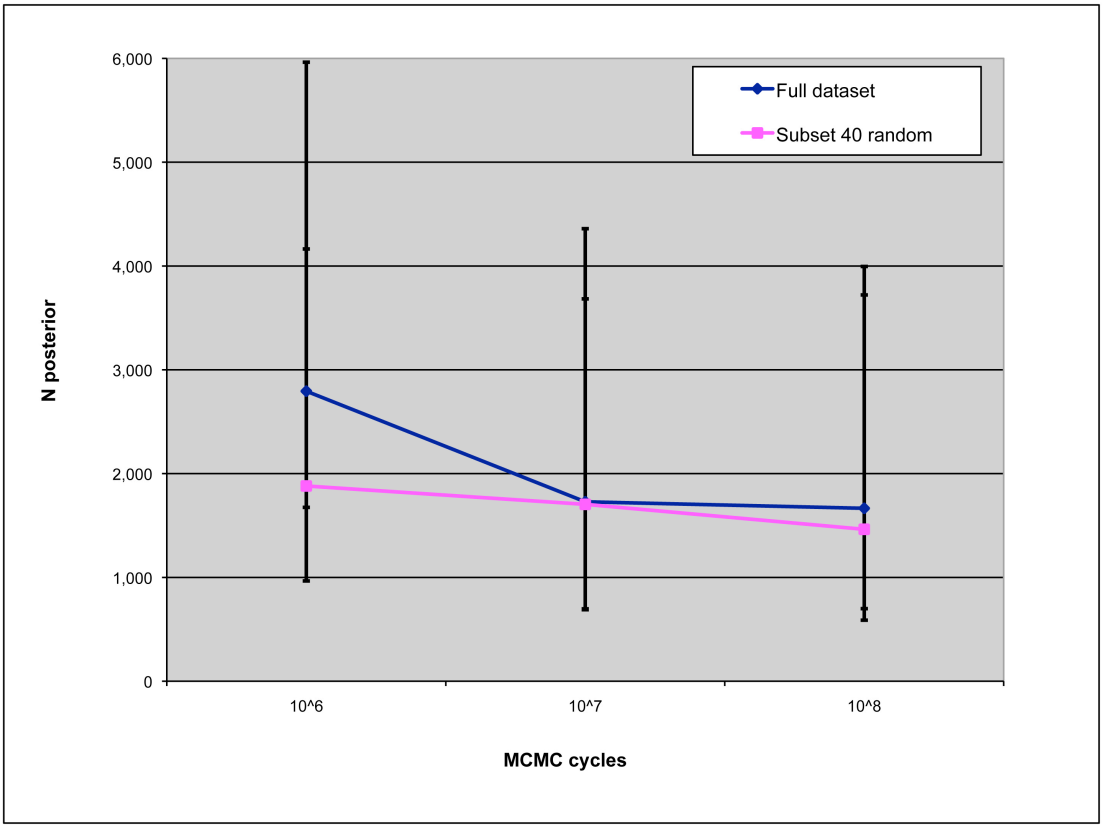


Figure B.4: Optimization of parameters for BATWING analysis. We performed BATWING analysis with a demographic model that assumes a period of constant size followed by exponential growth. To find the optimal number of MCMC cycles at which the method converges, we varied the MCMC cycles between 10^6 - 10^8 cycles and compared the results based on full dataset and a random subset of 40 samples. The number of MCMC cycles performed is shown on the X-axis and the posterior value of the effective population size (N) is shown on the Y-axis.

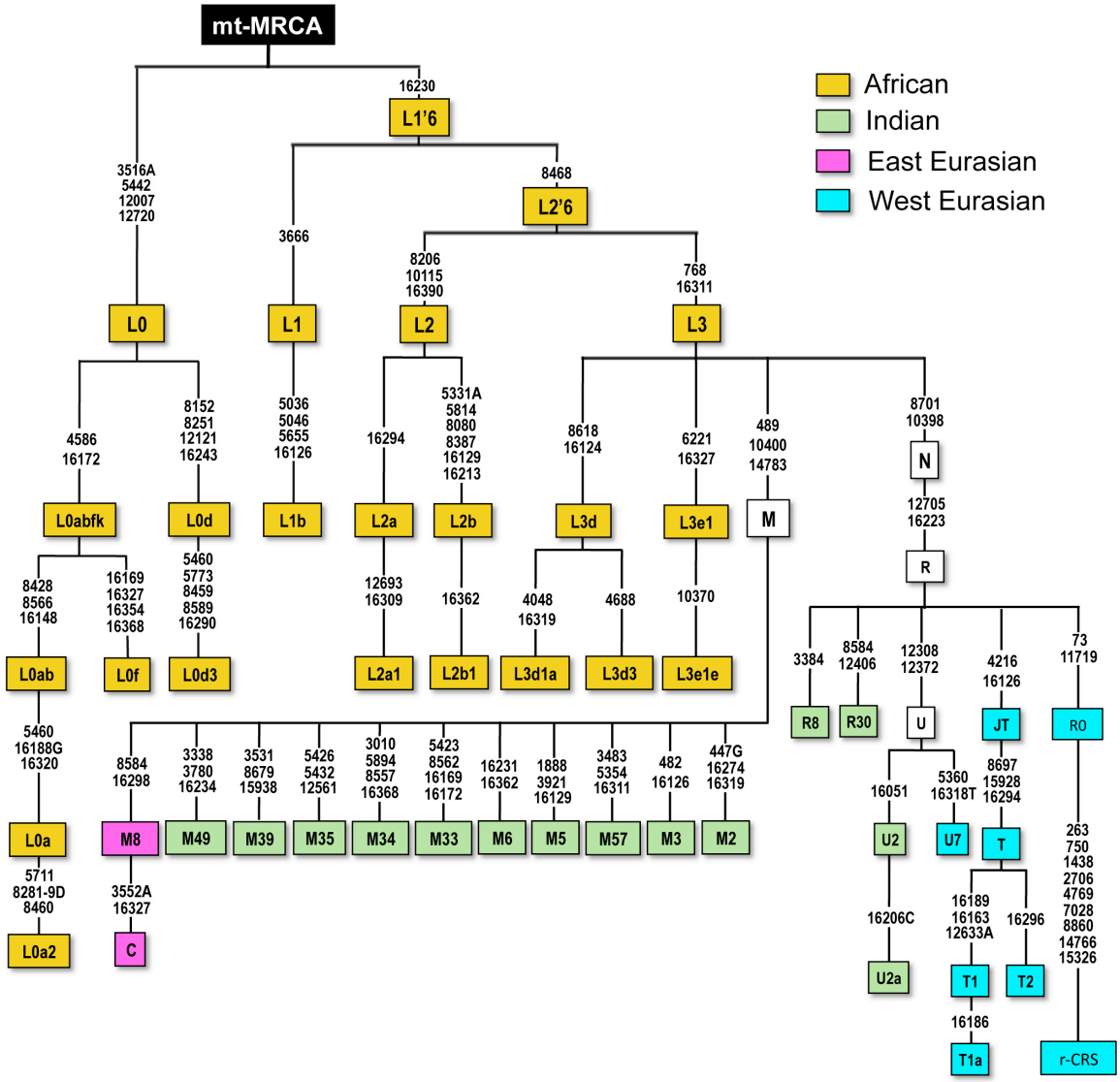


Figure B.5: Phylogenetic tree based on mitochondrial DNA markers. We performed mtDNA markers analysis for 153 Siddis and 269 individuals from the nearby Indian populations (Supporting Dataset S2). The following maternal lineages – African-specific L haplogroups, Indian-specific M, R and U lineages, East Asian branch of M and the Eurasian specific branches of R- were seen in the Siddi samples.

Table B.1: *F*st distribution in the present study

	Portugal	YRI	LWK	Siddi Karnataka1	Siddi Karnataka2	Siddi Gujarat	Mala	Madiga	Kurumba	Bhil	Kamsali	Satnami	Vysya	Naidu	Lodi	Tharu	Velama	Srivast-ava	Megha-wal	Vaish	Kashmiri Pandit	Hallaki
Portugal	0.00	0.15	0.14	0.09	0.10	0.09	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.05
YRI	0.15	0.00	0.01	0.04	0.03	0.03	0.15	0.15	0.15	0.15	0.15	0.15	0.16	0.15	0.15	0.14	0.15	0.14	0.14	0.14	0.15	0.15
LWK	0.14	0.01	0.00	0.04	0.02	0.03	0.14	0.14	0.13	0.13	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.14
Siddi Karnataka1	0.09	0.04	0.04	0.00	0.03	0.00	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.08	0.08
Siddi Karnataka2	0.10	0.03	0.02	0.03	0.00	0.02	0.09	0.09	0.09	0.09	0.09	0.09	0.10	0.09	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.09
Siddi Gujarat	0.09	0.03	0.03	0.00	0.02	0.00	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.09
Mala	0.05	0.15	0.14	0.08	0.09	0.08	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.02	0.01
Madiga	0.05	0.15	0.14	0.08	0.09	0.08	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.02	0.01
Kurumba	0.05	0.15	0.13	0.08	0.09	0.08	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.02	0.01
Bhil	0.05	0.15	0.13	0.08	0.09	0.08	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01
Kamsali	0.05	0.15	0.14	0.08	0.09	0.08	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01
Satnami	0.05	0.15	0.14	0.08	0.09	0.08	0.01	0.01	0.01	0.01	0.01	0.00	0.02	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.02	0.01
Vysya	0.06	0.16	0.14	0.09	0.10	0.09	0.01	0.01	0.01	0.01	0.01	0.02	0.00	0.02	0.01	0.01	0.02	0.01	0.02	0.02	0.02	0.02
Naidu	0.05	0.15	0.14	0.08	0.09	0.08	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Lodi	0.04	0.15	0.13	0.08	0.09	0.08	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
Tharu	0.04	0.14	0.13	0.08	0.08	0.08	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.01
Velama	0.04	0.15	0.13	0.08	0.09	0.08	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01
Srivastava	0.04	0.14	0.13	0.08	0.09	0.08	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.01
Meghawal	0.03	0.14	0.13	0.08	0.09	0.08	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01
Vaish	0.03	0.14	0.13	0.07	0.09	0.08	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01
Kashmiri Pandit	0.03	0.15	0.13	0.08	0.09	0.08	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.02
Hallaki	0.05	0.15	0.14	0.08	0.09	0.09	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.00

Note: Fst values were computed using EIGENSOFT

Table B.2: Formal test to confirm if Siddis have ancestry from Africans, Europeans and Indians

a. Is Ancestral pop2 necessary?				b. Is Ancestral pop2 sufficient?			
Admixed pop	**Ancestral pop1**	**Ancestral pop2**	**Z-score**	**Admixed**	**Ancestral pop1**	**Ancestral pop2**	**Z-score**
Siddi_Gujarat	Vaish	YRI	25.87	Siddi_Gujarat	YRI	Vaish	19.53
Siddi_Gujarat	Tharu	YRI	30.55	Siddi_Gujarat	YRI	Tharu	19.66
Siddi_Gujarat	Lodi	YRI	29.56	Siddi_Gujarat	YRI	Lodi	18.80
Siddi_Gujarat	Kashmiri Pandit	YRI	36.00	Siddi_Gujarat	YRI	Kashmiri Pandit	19.34
Siddi_Gujarat	Bhil	YRI	33.10	Siddi_Gujarat	YRI	Bhil	19.83
Siddi_Gujarat	Meghawal	YRI	30.52	Siddi_Gujarat	YRI	Meghawal	19.10
Siddi_Gujarat	Kurumba	YRI	32.34	Siddi_Gujarat	YRI	Kurumba	19.99
Siddi_Gujarat	Madiga	YRI	32.88	Siddi_Gujarat	YRI	Madiga	19.58
Siddi_Gujarat	Mala	YRI	31.91	Siddi_Gujarat	YRI	Mala	19.59
Siddi_Gujarat	Kamsali	YRI	29.58	Siddi_Gujarat	YRI	Kamsali	19.84
Siddi_Gujarat	Velama	YRI	28.81	Siddi_Gujarat	YRI	Velama	19.65
Siddi_Gujarat	Naidu	YRI	29.43	Siddi_Gujarat	YRI	Naidu	19.48
Siddi_Gujarat	Satnami	YRI	29.mai	Siddi_Gujarat	YRI	Satnami	19.40
Siddi_Gujarat	Hallaki	YRI	30.80	Siddi_Gujarat	YRI	Hallaki	18.75
Siddi_Gujarat	Vysya	YRI	33.77	Siddi_Gujarat	YRI	Vysya	19.02
Siddi_Gujarat	Portugal	YRI	39.26	Siddi_Gujarat	YRI	Portugal	19.22
Siddi_Gujarat	ICP	YRI	35.25	Siddi_Gujarat	YRI	ICP	18.87

NOTE: The Z-scores evaluates the significance of incorporating Ancestral pop2 in model - that is compares the model with ancestral pop2 against a null model with Ancestral pop2 excluded. Z-scores > |3| are statistically significant and implies that ancestral pop1 alone does not provide a good fit to the data. ICP - combined data from 16 Indian groups and Portuguese- represents the ancestral non-African population.

Table B.3: Identify the models that provide a good fit to the Siddi data

Admixed pop	pop1	pop2	pop3	Z-score
Siddi_Gujarat	YRI	Vaish	Portugal	**-1.45**
Siddi_Gujarat	YRI	Tharu	Portugal	**0.97**
Siddi_Gujarat	YRI	Kurumba	Portugal	4.22
Siddi_Gujarat	YRI	Lodi	Portugal	3.13
Siddi_Gujarat	YRI	Bhil	Portugal	3.59
Siddi_Gujarat	YRI	Madiga	Portugal	3.98
Siddi_Gujarat	YRI	Mala	Portugal	5.20
Siddi_Gujarat	YRI	Kamsali	Portugal	3.43
Siddi_Gujarat	YRI	Satnami	Portugal	3.37
Siddi_Gujarat	YRI	Meghawal	Portugal	**0.95**
Siddi_Gujarat	YRI	Pandit	Portugal	**0.29**
Siddi_Gujarat	YRI	Naidu	Portugal	**2.99**
Siddi_Gujarat	YRI	Velama	Portugal	**2.57**
Siddi_Gujarat	YRI	Vysya	Portugal	6.42
Siddi_Gujarat	YRI	Hallaki	Portugal	5.70
Siddi_Gujarat	YRI	cline	Portugal	**2.68**
Siddi_Gujarat	YRI	Portugal	Vaish	8.01
Siddi_Gujarat	YRI	Portugal	Tharu	8.61
Siddi_Gujarat	YRI	Portugal	Kurumba	8.98
Siddi_Gujarat	YRI	Portugal	Lodi	7.78
Siddi_Gujarat	YRI	Portugal	Bhil	9.40
Siddi_Gujarat	YRI	Portugal	Madiga	9.02
Siddi_Gujarat	YRI	Portugal	Mala	8.49
Siddi_Gujarat	YRI	Portugal	Kamsali	8.38
Siddi_Gujarat	YRI	Portugal	Satnami	7.92
Siddi_Gujarat	YRI	Portugal	Meghawal	6.99
Siddi_Gujarat	YRI	Portugal	Pandit	8.53
Siddi_Gujarat	YRI	Portugal	Naidu	8.12
Siddi_Gujarat	YRI	Portugal	Velama	7.48
Siddi_Gujarat	YRI	Portugal	Vysya	8.54
Siddi_Gujarat	YRI	Portugal	Hallaki	6.59
Siddi_Gujarat	YRI	Portugal	cline	9.90
Siddi_Gujarat	YRI	ICP	MEX	0.87

NOTE: The Z-scores evaluates the significance of incorporating Ancestral pop3 in model - that is compares the model with ancestral pop3 against a null model with Ancestral pop3 excluded. Highlighted in bold implies that the model of Ancestral Pop1 + Ancestral Pop2 provides a good fit to the data. "Indian cline" - combined data from 16 Indian groups- represents the ancestral Indian group. ICP - combined data from 16 Indian groups and Portuguese- represents the ancestral non-African population. HapMap3 Mexican ancestry in Los Angeles, California (MEX) are used to test if the YRI + ICP provides a good fit to the Siddi data as the variation in the Mexicans is unrelated to the genetics of Siddis.

Table B.4: Testing the robustness of the models that emerges from Table B.3

Admixed pop	Ancestral pop1	Ancestral pop2	Ancestral pop3	Z-score
Siddi_Karnataka-1				
Siddi_Karnataka-1	YRI	Vaish	Portugal	0.52
Siddi_Karnataka-1	YRI	Tharu	Portugal	-0.75
Siddi_Karnataka-1	YRI	Meghawal	Portugal	1.68
Siddi_Karnataka-1	YRI	Kashmiri Pandi	Portugal	0.62
Siddi_Karnataka-1	YRI	Naidu	Portugal	1.43
Siddi_Karnataka-1	YRI	Velama	Portugal	1.21
Siddi_Karnataka-1	YRI	Indian cline	Portugal	0.25
Siddi_Karnataka-1	YRI	ICP	MEX	-0.89
Siddi_Karnataka-2				
Siddi_Karnataka-2	YRI	Vaish	Portugal	-0.10
Siddi_Karnataka-2	YRI	Tharu	Portugal	-0.72
Siddi_Karnataka-2	YRI	Meghawal	Portugal	1.36
Siddi_Karnataka-2	YRI	Kashmiri Pandi	Portugal	-0.12
Siddi_Karnataka-2	YRI	Naidu	Portugal	0.51
Siddi_Karnataka-2	YRI	Velama	Portugal	1.07
Siddi_Karnataka-2	YRI	Indian cline	Portugal	1.72
Siddi_Karnataka-2	YRI	ICP	MEX	1.80

NOTE: ICP - combined data from 16 Indian groups and Portuguese- represents the ancestral non-African population and Indian cline - represents combined data for 16 Indian groups.

Table B.5: Testing the robustness of the models to the African population chosen for the analysis

Admixed	Ancestral pop1	Ancestral pop2	Ancestral pop3	Z-score
Siddi_Gujarat				
Siddi_Gujarat	LWK	Vaish	Portugal	1.07
Siddi_Gujarat	LWK	Tharu	Portugal	0.87
Siddi_Gujarat	LWK	Meghawal	Portugal	1.22
Siddi_Gujarat	LWK	Kashmiri Pandit	Portugal	1.08
Siddi_Gujarat	LWK	Naidu	Portugal	1.16
Siddi_Gujarat	LWK	Velama	Portugal	2.07
Siddi_Gujarat	LWK	Indian cline	Portugal	2.07
Siddi_Gujarat	LWK	ICP	MEX	-1.12

Table B.6: Result observed with the Batwing Analysis for determining the effective male population

MCMC cycles	N posterior				
	2.50%	Median	97.50%	Plus error	Minus error
Full dataset					
10^6	1,675	2,794	5,964	3,169	1,120
10^7	689	1,729	3,683	1,955	1,039
10^8	699	1,666	3,721	2,055	967
10^9					
Subset of 40 random chromosomes					
10^6	966	1,881	4,164	2,283	915
10^7	697	1,704	4,360	2,656	1,007
10^8	588	1,463	3,996	2,533	876
10^9					
Biased subset lacking haplogroup B					
10^6					
10^7	604	1,392	2,385		
10^8					
10^9					

NOTE: A random subset of 40 samples was analyzed using 10^6 to 10^8 MCMC cycles and we obtained the same posterior probability for effective population size (N) as that obtained for 10^7 cycles. We estimated the effective male population size of the African ancestors of Siddis brought to India as ~1,400 individuals.

Table B.7. *G6PD* variants observed in the Siddis and other Indian populations

Location				*G6PD* variant	
(State)	Population	Linguistic Family	N	A - variant rs1050828 rs1050829	Med variant rs5030868
Andhra Pradesh	Dudekula	Dravidian	18	-	-
	Erukala	Dravidian	39	-	-
	Patkar	Dravidian	11	-	-
	Thalari	Dravidian	12	-	-
	Vysya	Dravidian	25	-	-
Chattisgarh	Sindhi	Indo-European	91	-	5
Gujarat	Bhil	Indo-European	24	-	-
	Jat	Indo-European	8	-	-
	Siddi	Indo-European	60	6	1
Kerala	Kurumba	Dravidian	9	-	-
	Muduga	Dravidian	25	-	-
	Muduvar	Dravidian	62	-	-
Madhya Pradesh	Rajgond	Indo-European	24	-	-
	Sonr	Indo-European	72	-	-
Nagaland	Ao Naga	Tibeto-Burman	35	-	-
	Nagasema	Tibeto-Burman	38	-	-
	Chakhesang Naga	Tibeto-Burman	16	-	-
Uttar Pradesh	Yadav	Indo-European	22	-	-
	Lodhi	Indo-European	47	-	-
Uttaranchal	Bisth	Indo-European	36	-	-
	Uniyal	Indo-European	34	-	-
Karnataka	Siddi	Dravidian	67	7	-
	Kare Vokkal	Dravidian	30	-	-
	GramVokkal	Dravidian	54	-	-
	Korava	Dravidian	38	-	-
	Medar	Dravidian	56	-	-
Tamil Nadu	Khani	Dravidian	46	-	-
	Badaga	Dravidian	57	-	-
Total			**1056**		

Appendix C

Supplementary Material for Chapter 4

Note C.1: New *ROLLOFF* statistic.

In this note, we consider alternative forms of the *ROLLOFF* linkage disequilibrium (LD) statistic[1] for dating population admixture events. We show that the original *ROLLOFF* statistic is susceptible to downward bias in the event of a recent population bottleneck, and we propose a modification of the statistic that is robust against such an effect (Table C.3).

The *ROLLOFF* technique applies two key insights: first, that admixture creates LD that decays exponentially as recombination occurs – explicitly, as e^{-nd} where n is the number of generations since admixture and d is the genetic distance between SNPs – and second, that the amount of admixture LD between each pair of SNPs is proportional to the product of the allele frequency divergences between the ancestral populations at those sites. The latter observation allows the e^{-nd} admixture LD decay signal to be detected (via a SNP-pair weighting scheme) and harnessed to infer the mixture date n.

The original *ROLLOFF* statistic captures admixture LD in the form of SNP autocorrelation. Defining $z(x,y)$ to be the (Fisher z-transformed) correlation coefficient between SNP genotypes at sites x and y, *ROLLOFF* computes the correlation coefficient between values of $z(x,y)$ and weights $w(x,y)$ over pairs of SNPs binned by genetic distance:

$$A(d) := \frac{\sum_{|x-y| \approx d} z(x,y) w(x,y)}{\sqrt{\sum_{|x-y| \approx d} z(x,y)^2} \sqrt{\sum_{|x-y| \approx d} w(x,y)^2}} \qquad (1)$$

the idea being that $A(d) \propto e^{-nd}$.

While this setup estimates accurate dates for typical admixture scenarios, it turns out to be noticeably biased in the case of a recent bottleneck. However, we will show that the following modified statistic does not suffer from the bias:

$$R(d) := \frac{\sum_{|x-y|\approx d} z(x,y)w(x,y)}{\sum_{|x-y|\approx d} w(x,y)^2} \qquad (2)$$

(Note that *R(d)* amounts to taking the regression coefficient of *z(x,y)* against the weights *w(x,y)* for SNP pairs within each bin).

An additional detail of our modified *ROLLOFF* statistic is that we change *z(x,y)* to measure admixture LD as the covariance between SNPs *x* and *y* rather than the correlation (i.e., it equals the classical LD statistic *D* rather than the correlation *r*). We believe the use of covariance rather than correlation for *z(x,y)* has little impact on the performance and properties of the statistic (as it roughly amounts to multiplying by a constant factor) but makes the statistic more amenable to mathematical analysis.

Explanation of bias from recent bottlenecks

The bias in the original formulation of *ROLLOFF* (1) introduced by a recent bottleneck can be readily explained at an intuitive level: the problem is that while the numerator of the correlation coefficient, $\sum_{|x-y|\approx d} z(x,y)w(x,y)$, decays as e^{-nd} as intended, the normalization term:

$$\sqrt{\sum_{|x-y|\approx d} z(x,y)^2} \qquad (3)$$

Also exhibits a decay behavior that confounds the e^{-nd} signal (Figure C.3). The reason is that a strong bottleneck introduces a very large amount of LD, effectively giving *z(x,y)* a random large magnitude immediately post-bottleneck that is independent of the distance between *x* and *y*. This LD subsequently decays as e^{-nd} until the magnitude of *z(x,y)* reaches the level of random sampling noise (arising from the finite sample of admixed individuals being used to calculate *z*). In non-bottlenecked cases, the square-norm of *z(x,y)* is usually dominated by sampling noise, so the normalization term (3) effectively amounts to a constant, and dividing by (3) has no effect on the decay rate of *A(d)*.

The "regression coefficient" version of the *ROLLOFF* statistic (2) does not contain the normalization term (3) and thus does not incur bias from bottlenecks.

Precise effect of genetic drift on the *ROLLOFF* statistics

We now rigorously derive the above intuition. We will assume in the following calculations that the *ROLLOFF* weights are taken as the product of allele frequency divergences $\delta(x)$ and $\delta(y)$ in the ancestral mixing populations:

$$w(x,y) := \delta(x)\delta(y)$$

Our reasoning below applies whether we have the true values of $\delta(x)$ and $\delta(y)$ or computed weights based on related reference populations or PCA loadings. We also assume that all SNPs are polymorphic ancestrally---i.e., we ignore mutations that have arisen in the admixed population---and that the SNP ascertainment is unbiased with respect to the populations under consideration.

For a diploid population of size N with chromosomes indexed by $i = 1, \ldots, 2N$, we set

$$z(x,y) := \frac{1}{2N}\sum_{i=1}^{2N}(X_i - \mu_x)(Y_i - \mu_y)$$

to be the covariance between binary alleles X_i and Y_i at sites x and y, respectively. We assume for ease of discussion that the data are phased; for unphased data, *z(x,y)* is essentially a noisier version of the above because of cross terms.

We are primarily interested in the behavior of *z(x,y)* from one generation to the next. Fix a pair of SNPs x and y at distance d and let z_0 denote the value of *z(x,y)* at a certain point in time. After one generation, due to finite population size and recombination, the covariance becomes[2]

$$z_1 = z_0 e^{-d}(1 - 1/2N) + \varepsilon, \qquad (4)$$

where N is the population size, e^{-d} is the probability of no recombination, (*1-1/2N*) is a Bessel correction, and ε is a noise term with mean 0 and variance on the order of *1/N*.

Iterating this equation over n generations, the final covariance is

$$z_n = z_0 e^{-nd} e^{-n/2Ne} + \varepsilon_{agg}$$

where N_e is the effective population size over the interval and ε_{agg} is a sum of n partially decayed noise terms.

Now let time 0 denote the time of admixture between two ancestral populations mixing in proportions α and $\beta := 1-\alpha$. (The bottleneck may have occurred either before or after this point, as long as it does not influence the calculation of the weights). Then $E[z_0]$ is:

$$E[z_0] = 2\alpha\beta\delta(x)\delta(y)$$

Assuming the mixture is homogeneous and the distance d is large enough that background LD can be ignored. (In practice, heterogeneity in the admixed population changes the above form and results in the addition of an affine term to the *ROLLOFF* curve, which we explicitly fit. We also typically fit only data from SNP pairs at distance $d > 0.5$ cM to avoid background LD).
We can now compute the modified *ROLLOFF* statistic:

$$E[R(d)] = E\left[\frac{\sum_{|x-y|\approx d} z(x,y)\delta(x)\delta(y)}{\sum_{|x-y|\approx d} \delta(x)^2\delta(y)^2}\right]$$

$$\approx \frac{\sum_{|x-y|\approx d} [2\alpha\beta\delta(x)\delta(y)e^{-nd}e^{-n/2Ne} + \varepsilon_{agg}]\delta(x)\delta(y)}{\sum_{|x-y|\approx d} \delta(x)^2\delta(y)^2}$$

$$\approx 2\alpha\beta e^{-nd}e^{-n/2Ne}$$

Importantly, in the last step we use the fact that the combined noise term ε_{agg} is uncorrelated with $\delta(x)\delta(y)$. Thus, even a strong bottleneck with a low value of N_e only scales $R(d)$ by the constant factor $e^{-n/2Ne}$, and the e^{-nd} scaling of the *ROLLOFF* curve as a function of d is unaffected.

On the other hand, if we use the original correlation form (1) of the *ROLLOFF* statistic *A(d)*, then the numerator still has the form of an exponential decay Ae^{-nd}, but now we divide this by the norm $\sqrt{\sum_{|x-y|\approx d} z(x,y)^2}$. In the case of a strong bottleneck, $z(x,y) = z_0 e^{-nd} e^{-n/2Ne} + \varepsilon_{agg}$ can be dominated by the aggregate noise term ε_{agg}. Indeed, if the bottleneck occurred *k* generations ago, then the noise terms ε_i from the time of reduced population size will have decayed by e^{-kd} since the bottleneck but can still have large variance if the population size N_{bot} was very small at the time. In this case, at lower values of *d*, $E[z(x,y)^2] = E[(z_0 e^{-nd} e^{-n/2Ne} + \varepsilon_{agg})^2]$ will be dominated by $E[\varepsilon_{agg}^2]$ which will scale approximately as e^{-2kd}/N_{bot}. Hence, the denominator of *A(d)* will be significantly larger at low *d* than at high *d*, causing a partial cancellation of the exponential decay of the *ROLLOFF* curve and thus a downward bias in the estimated date of admixture.

Note C.2: Simulations for estimating dates of admixture events.

Simulation 1: To test the effect of founder events post admixture

In order to test the effect of founder events post admixture, we performed simulations using MaCS[3] coalescent simulator. We simulated data for three populations (say, *A*, *B* and *C*). We set the effective population size (N_e) for all populations to 12,500 (at all times except during the founder event), mutation and recombination rate were set $2\text{x}10^{-8}$ and to $1\text{x}10^{-8}$ per base pair per generation respectively. *C* can be considered as an admixed population that has 60%/40% ancestry from *A'* and *B'* (admixture time (*t*) was set to 30/ 100 generations before present). *A'* and *A* diverged 120 generations ago, *B'* and *B* diverged 200 generations ago and *A* and *B* diverged 1800 generations ago. At generation *x* ($x < t$), *C* undergoes a severe founder event

where the effective population size (N_e) reduces to 5 individuals for one generation. At generation ($x+1$), the N_e = 12,500. We simulated data for 5 replicates for each parameter. We performed *ROLLOFF* analysis (using the original and modified statistic) with *C* as the target and *A* and *B* as the reference populations. When we use the original *ROLLOFF* statistic (*A(d)*), we observe that the dates are biased downward in cases of founder events post admixture. However, when we use the modified statistic (*R(d)*), the bias is removed (Table C.3). Details of the bias correction are shown in Note C1. Throughout the manuscript, we use the modified *ROLLOFF* statistic (*R(d)*) unless specified otherwise.

Simulation 2: To test the accuracy of the modified *ROLLOFF* statistic (*R(d)*)

We perform simulations using the same simulation framework as in reference 1 to test the accuracy of the estimated dates using the modified *ROLLOFF* statistic. We simulated data for 25 admixed individuals using Europeans (HapMap CEU) and HGDP East Asians (Han) as ancestral populations, where mixture occurred between 10-300 generations ago and European ancestry proportion was set to 20%. These ancestral populations were chosen as F_{st}(CEU, Han) = 0.09 is similar to the F_{st} between the ancestral populations of the Roma (Europeans and ASI). Figure C.4 shows that we get accurate estimates for the dates of mixture up to 300 generations.

Simulation 3: To test the effect of using PCA loadings instead of allele frequencies as weights in *ROLLOFF*

In the case of Roma admixture, data from unadmixed South Asian populations is not available and so it is not possible to compute the allele frequencies of SNPs for one of ancestral populations (ASI). However, data from many South Asian populations (which are admixed with

ANI and ASI ancestry) are available and can be used for estimating the PCA-based SNP loadings. We performed simulations to mimic this scenario.

We simulated data for 60 admixed individuals using Europeans (HapMap CEU) and HGDP East Asians (Han) as ancestral populations, where mixture occurred 100 generations ago and European ancestry proportion was set to 30% (group 1: n = 20), 50% (group 2: n = 20) and 70% (group 3: n = 20). These three groups of simulated samples can be roughly considered as three South Asian populations. We performed PCA analysis with CEU and Groups 1-3 of simulated samples to estimate the SNP loadings that can be used in *ROLLOFF*.

Next, we simulated data for 54 individuals that can be used as the target in the *ROLLOFF* analysis. These individuals have 80%/20% European and East Asian ancestry respectively (similar to Roma) and the date of mixture is set to 30 (n = 27) and 100 (n = 27) generations before present. We ran *ROLLOFF* (using *R(d)*) to estimate the date of mixture in this panel of individuals using the PCA-based loadings computed above. We estimated that the dates of mixture were 33 ± 1 and 99 ± 1 generation for mixture that occurred 30 and 100 generations ago respectively (Figure C.5). This shows that we can effectively estimate the date of mixture even in the absence of data from unadmixed ancestral populations, as long as data from other admixed individuals (involving the relevant ancestral populations) is available.

Simulation 4: To test the model of two waves of admixture

In order to obtain an interpretation of the *ROLLOFF* estimated date of mixture when the assumption of single wave of mixture is incorrect, we ran *ROLLOFF* (using *R(d)*) to infer the date of admixture for data simulated under a two pulse admixture scenario. We simulated data using Europeans (HapMap CEU) and HGDP East Asians (Han) as the ancestral populations

using the simulation framework described in reference 1. We simulated two pulse admixture scenarios in which a 50%/50% admixture of CEU and Han occurred at λ_1, followed by a 60%/40% mixture of that admixed population and CEU at λ_2 (Table C.4). The mixture proportions were chosen so that the final European ancestry proportion is ~80% (similar to Roma). We ran *ROLLOFF* (using *R(d)*) with a non-overlapping set of Europeans and Han as the reference population. Table C.4 shows that as the interval between the time of the gene flow events (λ_2-λ_1) increases, the estimated dates of mixture reflects the date of the more recent gene flow event.

Note C.3: Computing corrected IBD sharing distance between Roma and South Asian groups.

To find the source of the South Asian ancestry in Roma, we inferred the pairwise IBD sharing distance between Roma and various South Asian groups using GERMLINE[4]. We observed that the Roma share the highest proportion of IBD sharing with groups from the northwest of India (Figure 4.3b). We were concerned that high IBD sharing could be an artifact related to the high proportion of ANI ancestry in the North-western Indian groups. Hence, we performed a regression analysis to correct for the effect of the ANI ancestry proportion on IBD sharing distance. The model that provided the best fit was IBD sharing = 0.35 + 0.81*ANI ancestry proportion (P-value < 0.05). Each South Asian group was considered as a single data point for this analysis. Next, we computed an average corrected IBD sharing measure for each region by regression out the effect of ANI ancestry and computing an average of the residuals for each region in India. Note: For this analysis, we did not include the Eastern Indian populations

(Nysha and Ao Naga) and Andamanese populations (Onge and Great Andamanese) as these populations are not simple admixtures of ANI and ASI groups.

In order to control for the effect of the sample size on the IBD computation, we performed bootstrap analysis such that for each run, we randomly sampled up to 30 individuals (some groups had < 30 samples) from each of the 8 regional groups and estimated the IBD sharing statistics between Roma and the regional group. We performed a total of 100 runs and obtained the mean and standard error of the IBD statistic (Figure C.7). We observed that Roma still share the highest proportion of IBD segments with groups from Northwest of India.

Note C.4: Simulations for estimating date of founder event.

We used MaCS[3] coalescent simulator to perform simulations to test the robustness of our allele sharing statistic that we use for estimating the dates of the founder event. We simulate data for two populations (say, *A* and *B*) that diverged 1800 generations ago. We set the effective population size for both populations as N_e = 12,500, mutation rate = $2x10^{-8}$ and recombination rate = $1x10^{-8}$ per base pair per generation respectively. For each simulation, we compute the autocorrelation of allele sharing within individuals of *B*, and then subtract the cross-population autocorrelation between *A* and *B* to remove the effects of ancestral allele sharing (see Methods).

Simulation 1: Founder event only

B undergoes a severe founder event *x* generations ago where the effective population size reduces to 5 individuals for one generation. At generation (*x+1*), the population size = N_e again. Table C.5 shows that in such cases we can accurately estimate the date of the founder event using our statistic.

Simulation 2: Founder event and admixture

We simulate data for a more complex demography where *B* is admixed and has 40% ancestry from *A'* which is closely related to *A*. The admixture occurred at time *t* and at time *x* = 10, 30 or 100 generations, *B* undergoes a severe founder event where the effective population size of *B* reduces to 5 individuals for one generation. Table C.5 shows that for a recent founder event (10 and 30 generations ago), we accurately estimate the date of the founder event. However, for older founder events (100 generations), we are unable to accurately estimate the date of the founder event, if it occurred pre-admixture. However, this is expected as we are only sampling the admixed population (today) and not the ancestral population that underwent the founder event.

Simulation 3: No Founder event

We simulate data for a complex demography where *B* is admixed and has 40% ancestry from *A'* which is closely related to *A*. The admixture occurred 10, 30, 50 or 70 generations ago. In all cases, we observe that the allele-sharing statistic is not associated to distance. We test if the model of a straight line ($y \sim c$) or exponential decay ($y \sim c + Ae^{-tD}$), where *D* = genetic distance and *t* = time of founder event) provides a better fit to the output. In all four cases, we fail to reject the null model ($y \sim c$) (P > 0.05).

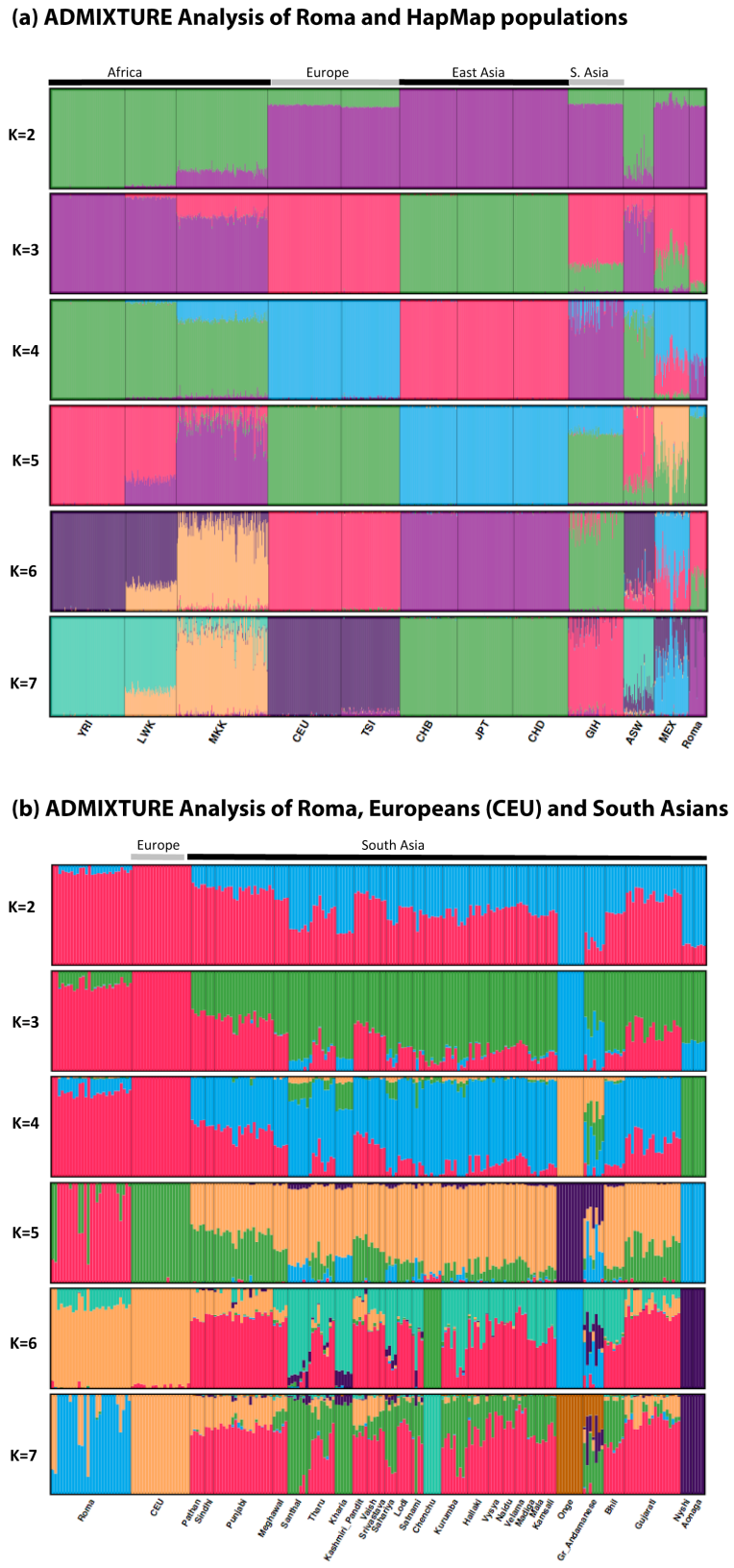


Figure C.1: ADMIXTURE Analysis.

Figure C.1 (Continued) To study the relationship of Roma with worldwide populations, we performed ADMIXTURE analysis. Each vertical line represents an individual colored based on the proportion of estimated ancestry for each cluster. (a) ADMIXTURE Analysis (K=2 to K=7) of Roma and HapMap populations. Lowest cross validation error was observed for K=6; (b) ADMIXTURE Analysis of Roma, Europeans (CEU) and South Asians. Lowest cross validation error was observed for K=3. We limit the sample size of all groups (except Roma) to 20 individuals.

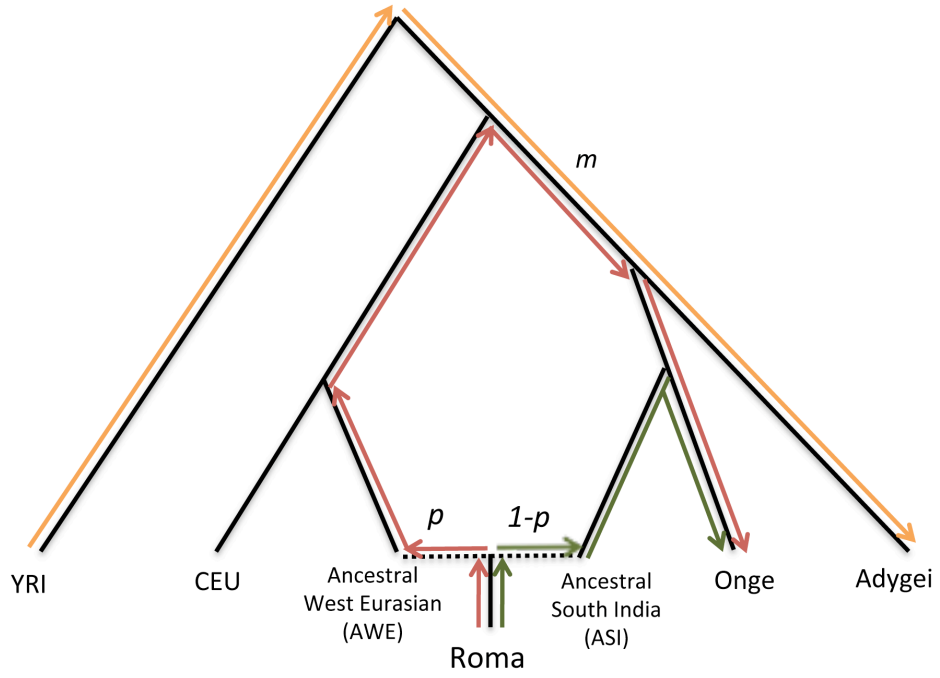
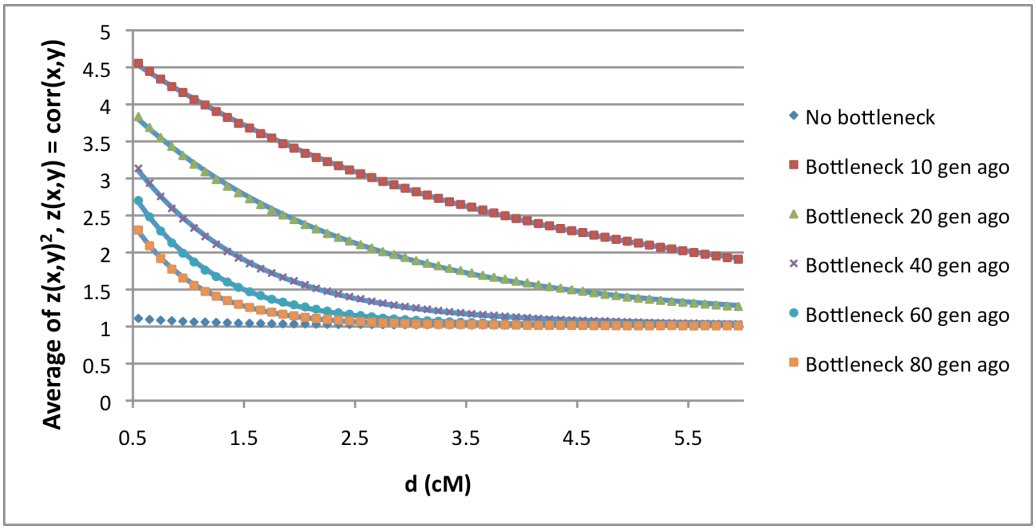


Figure C.2: Estimating the proportion of West Eurasian and South Asian ancestry in Roma. In order to estimate the proportion of West Eurasian ancestry in Roma, we use the phylogenetic tree shown below. The different colored lines show drift that has occurred between the populations connected by the line. The orange line shows the drift between YRI and Adygei and the red and green lines shows the drift separating Roma and Onge. *m* denotes the shared drift between Roma and Onge. See methods for details for estimating the West Eurasian ancestry proportion (*p*) in Roma that derives from India (ANI) and Europe (post exodus from India). This figure is adapted from Reich et al (2009).

(a) Using $z(x,y)$ = correlation(x,y)



(b) Using $z(x,y)$ = covariance(x,y)

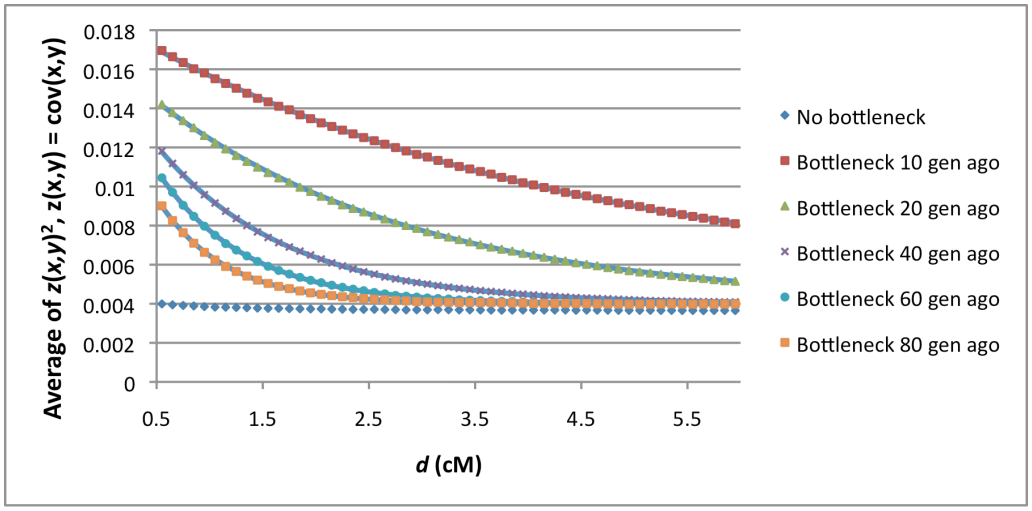


Figure C.3: Normalization term from original *ROLLOFF* correlation coefficient formulation. We plot the squared normalization term $\sum z(x,y)^2$ as a function of genetic distance d between SNPs for the admixture plus bottleneck scenarios described in Table C.3, using either the correlation (a) or covariance (b) versions of $z(x,y)$. In the case of no bottleneck, the normalization term is dominated by finite sampling noise and exhibits no dependence on d. For the cases of a strong bottleneck post-admixture, however, $\sum z(x,y)^2$ exhibits an exponential decay $Ae^{-2kd} + c$ with rate constant approximately equal to twice the age of the bottleneck ((a) best-fit k = 15, 25, 46, 65, 83 and (b) k = 12, 20, 41, 60, 78 shown as solid lines).

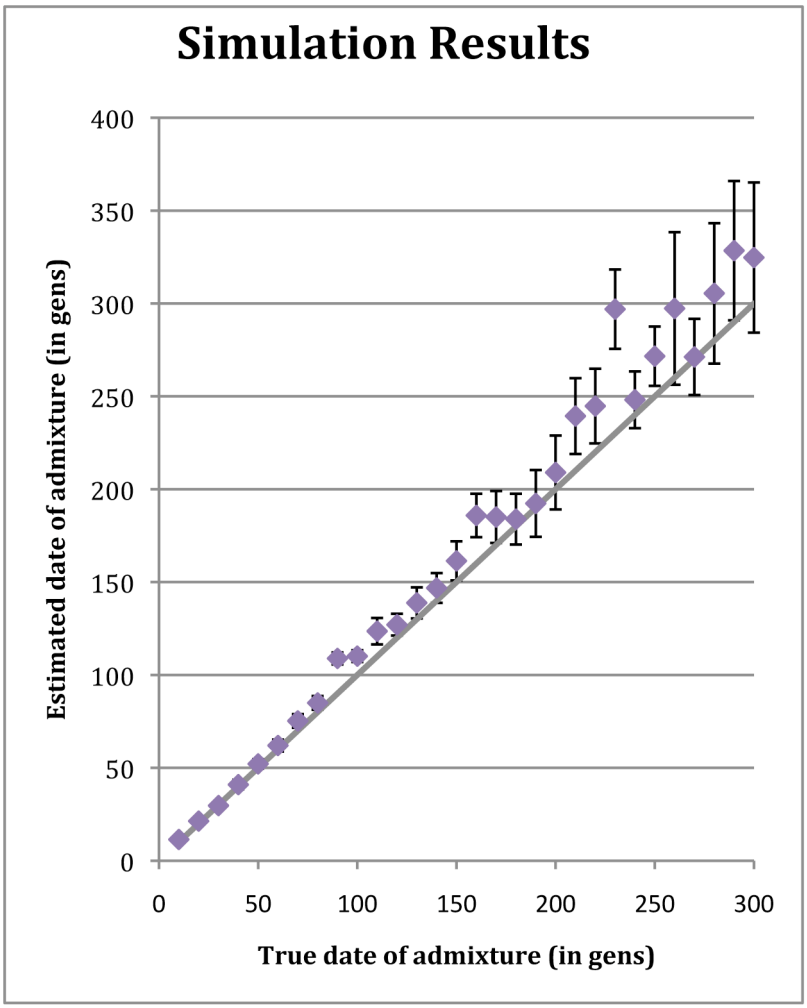


Figure C.4: *ROLLOFF* Simulation Results: Variable age of mixture. We simulated data for 25 admixed individuals with mixed European and East Asian ancestry where the proportion of European ancestry was set to 20% and the admixture date was set between 10-300 generations (as shown below). We ran the *ROLLOFF* (using *R(d)*) to estimate the date of mixture using allele frequencies in an independent dataset of French and East Asians. Standard errors were computed using weighted block jackknife as described in the Methods.

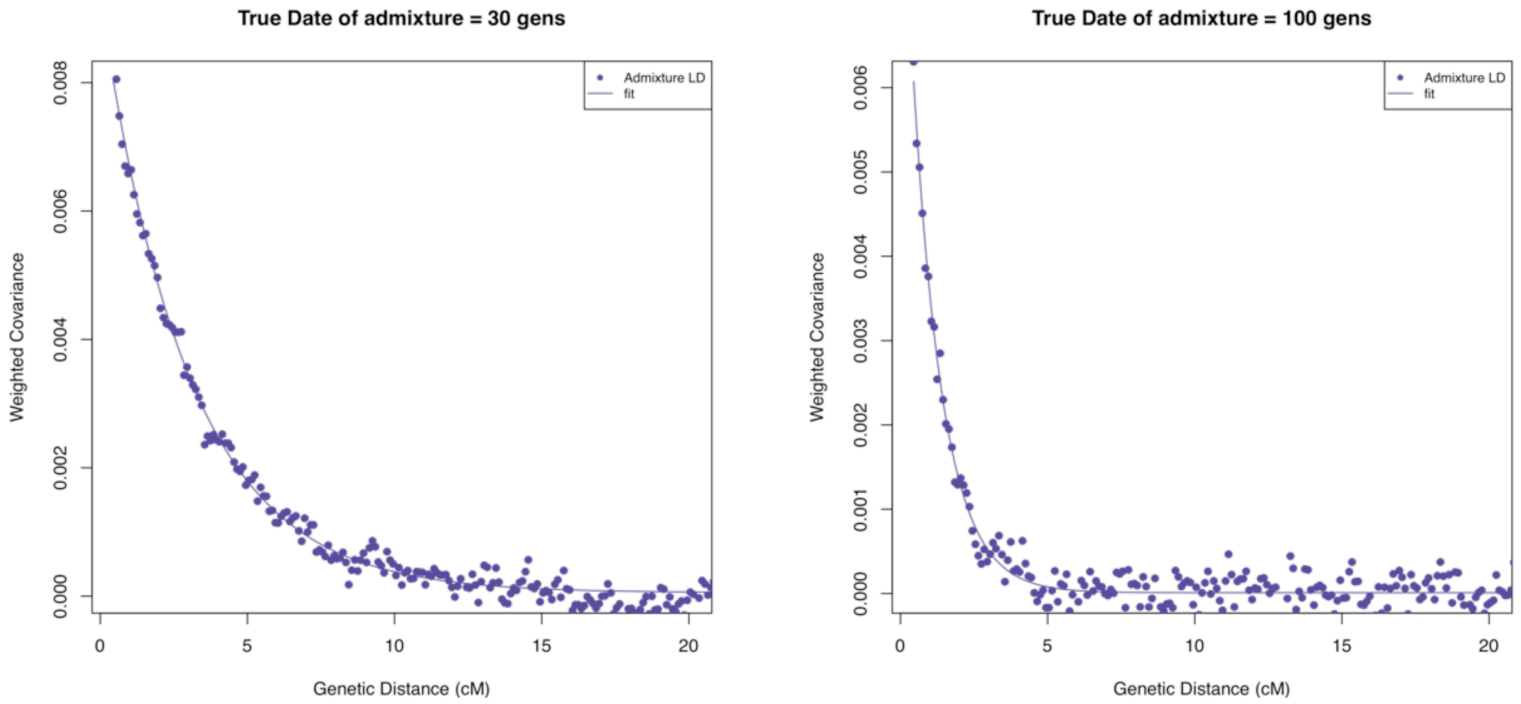


Figure C.5: *ROLLOFF* Simulation using PCA-loadings. We simulated data for admixed individuals with mixed European and East Asian ancestry where the proportion of European ancestry was set to 80% (similar to Roma) and the mixture occurred 30 generations ago (left panel: n = 27) and 100 generations ago (right panel: n = 27). We ran *ROLLOFF* (using *R(d)*) to estimate the date of mixture in this panel of individuals using the PCA-based loadings computed using an CEU and an independent dataset containing simulated data for 3 admixed groups with European ancestry equal to 30%, 50% and 70%. We estimated that the dates of mixture were 33 ± 1 generation for the left panel (true date = 30 gens), and 99 ± 4 generations for right panel (true date = 100 gens).

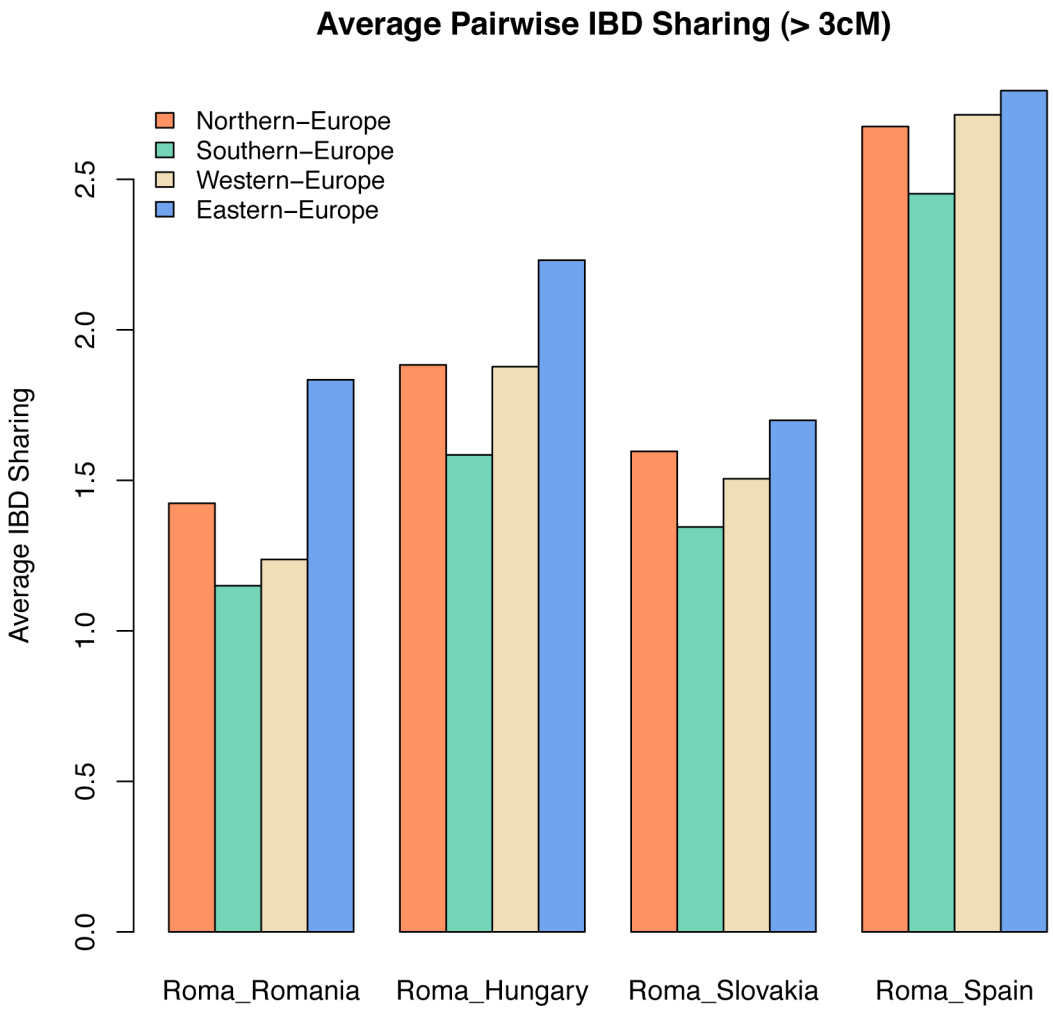


Figure C.6: IBD Sharing of Roma with European populations. We computed average pairwise IBD sharing between Roma from European samples (from POPRES, HapMap and HGDP datasets) clustered based on geography.

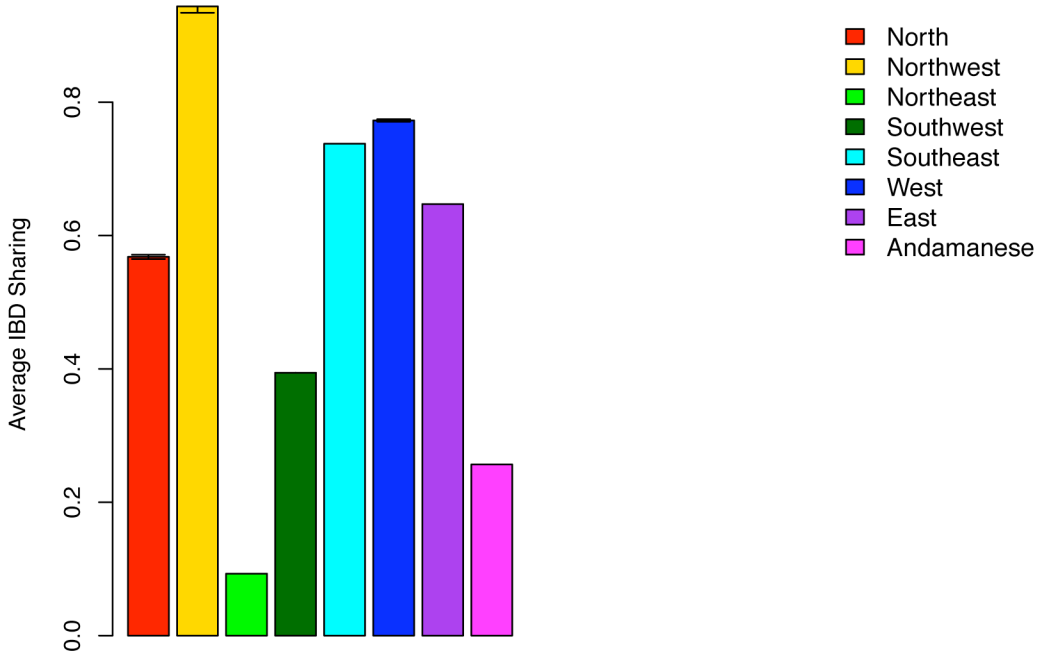


Figure C.7: Bootstrap analysis to compute error in IBD statistics. We performed bootstrap analysis where we randomly sample up to 30 individuals from each of the 8 South Asian regional groups and compute average pairwise IBD between Roma and South Asians. We performed a total of 100 runs and obtained the mean and standard error for the IBD statistic (vertical bars shown). For regional groups, which had less than 30 samples (such as Northeast, Southwest, East, and Andamanese), all samples were included in each run and so no standard errors are shown.

Table C.1: Average frequency differentiation (F_{st}) for Roma and HapMap populations

	CEU	YRI	CHB	JPT	ASW	CHD	GIH	LWK	MEX	MKK	TSI	Roma
CEU	0	0.14	0.102	0.104	0.088	0.103	0.033	0.13	0.036	0.093	0.003	0.016
YRI	0.14	0	0.169	0.17	0.008	0.169	0.129	0.007	0.134	0.025	0.136	0.135
CHB	0.102	0.169	0	0.007	0.127	0.001	0.071	0.159	0.064	0.131	0.102	0.092
JPT	0.104	0.17	0.007	0	0.129	0.008	0.072	0.161	0.065	0.133	0.104	0.094
ASW	0.088	0.008	0.127	0.129	0	0.128	0.083	0.009	0.088	0.013	0.086	0.087
CHD	0.103	0.169	0.001	0.008	0.128	0	0.071	0.16	0.066	0.132	0.103	0.093
GIH	0.033	0.129	0.071	0.072	0.083	0.071	0	0.119	0.038	0.086	0.032	0.026
LWK	0.13	0.007	0.159	0.161	0.009	0.16	0.119	0	0.125	0.015	0.126	0.125
MEX	0.036	0.134	0.064	0.065	0.088	0.066	0.038	0.125	0	0.093	0.037	0.04
MKK	0.093	0.025	0.131	0.133	0.013	0.132	0.086	0.015	0.093	0	0.088	0.089
TSI	0.003	0.136	0.102	0.104	0.086	0.103	0.032	0.126	0.037	0.088	0	0.015
Roma	0.016	0.135	0.092	0.094	0.087	0.093	0.026	0.125	0.04	0.089	0.015	0

Table C.2: Formal tests of admixture

Population (*X*)	Sam-ples	Region	*Z*-score for *4 Population test* $(P_{CEU}-P_{YRI}) \times (P_{Onge}-P_X)$	$(P_{YRI}-P_{Onge}) \times (P_{CEU}-P_X)$	$(P_x-P_{YRI}) \times (P_{CEU}-P_{Onge})$	Estimated West Eurasian Ancestry %
Roma	18	Hungary	-33	4.8	-29.3	78.3 ± 1.9%
Roma*	3	Slovakia	-26.6	3.5	-22.8	71.5 ± 3.1%
Roma**	1	Romania	-20.2	0.7	-19.2	79.4 ± 4.7%
Roma	2	Spain	-25.3	0.9	-24	75.6 ± 4.0%
Roma	24	Combined	-33	4.8	-29.5	77.5 ± 1.8%

NOTE: * indicates that some samples from the group appear to have recent European gene flow. These samples were excluded from the analysis (the number of * indicates the number of samples excluded). Ancestry proportions were estimates based on f_4 *Ratio Estimation* using Yoruba, Adygei, Europeans (CEU) and Onge as the reference populations.

Table C.3: Simulations for estimating dates of admixture events: Founder events post admixture model

True date of admixture	True date of founder event (*x*)	Date based on original *ROLLOFF* statistic (a)	Date based on modified *ROLLOFF* statistic (b)	Date based on modified *ROLLOFF* statistic (c)
30	N/A	31.3	32.0	32.1
30	5	24.6	30.1	29.0
30	10	27.7	34.1	32.3
30	20	23.3	32.7	31.0
30	25	23.4	30.8	29.5
100	N/A	94.1	96.8	97.0
100	10	93.9	106.1	102.9
100	20	87.1	102.7	97.3
100	40	75.3	95.6	92.2
100	60	83.9	106.3	102.8
100	100	81.6	101.1	99.0

Note: We simulated data from three populations Pop A (n = 20), Pop B (n = 20) and Pop C (n = 30) using MaCS coalescent simulator. Populations A and B diverged 1800 generations ago. The effective population size for all populations was set 12,500 at all times (except during the founder event). The mutation and recombination rates were set to $2x10^{-8}$ and $1x10^{-8}$ per base pair per generation. Pop C can be considered as an admixed population that has ancestry 60%/40% ancestry from A' and B' (admixture time (t) is set to 30/ 100 generations). Pop A' and A diverged 120 generations and B' and B diverged 200 generations ago. At generation *x* (shown in table above), Pop C undergoes a severe founder event where the effective population size reduces to 5 individuals for one generation. When *x* = N/A, there was no founder event. We performed *ROLLOFF* (using original and modified statistic) with Pop C as the target and Pop A and B as the reference populations. We performed 5 replicates for each parameter and report the average estimated date of mixture. The statistics used were -

(a) Original *ROLLOFF* Statistic: $A(d)=\frac{\sum_{|x-y|\approx d} z(x,y)w(x,y)}{\sqrt{\sum_{|x-y|\approx d} z(x,y)^2}\sqrt{\sum_{|x-y|\approx d} w(x,y)^2}}$; where $z(x,y)$ = correlation between *x* and *y*.

(b) Modified Statistic: $R(d)=\frac{\sum_{|x-y|\approx d} z(x,y)w(x,y)}{\sum_{|x-y|\approx d} w(x,y)^2}$; where $z(x,y)$ = correlation between *x* and *y*.

(c) Modified Statistic: $R(d)=\frac{\sum_{|x-y|\approx d} z(x,y)w(x,y)}{\sum_{|x-y|\approx d} w(x,y)^2}$; where $z(x,y)$ = ***covariance*** between *x* and *y*.

Table C.4: Simulations for estimating dates of admixture events: Model with two gene flow events

Date of first wave of mixture (λ_1)	Date of second wave of mixture (λ_2)	Estimated date in generations (± standard error)
120	20	36 ± 3
170	20	28 ± 2
220	20	23 ± 2
270	20	24 ± 2
320	20	25 ± 1
370	20	25 ± 1
420	20	22 ± 1
130	30	46 ± 3
180	30	47 ± 3
230	30	41 ± 2
280	30	39 ± 2
330	30	39 ± 3
380	30	35 ± 2
430	30	32 ± 3

Note: We simulated 27 individuals using CEU and Han Chinese as the ancestral populations where we set the overall European ancestry proportion to be 80%. We then performed *ROLLOFF* (using *R(d))* with an independent dataset of Europeans (HGDP French) and East Asians (HapMap CHB) as reference populations.

Table C.5: Simulations for estimating dates of founder events

Simulation scenario	True date of founder event	True date of admixture	Estimated date of founder event (in generations)
Founder event only			
	10	--	11.2
	20	--	20.8
	40	--	39.3
	60	--	52.7
	80	--	74.9
	100	--	95.7
Founder event + Admixture			
	10	10	8.2
	10	20	8.4
	10	40	8.3
	10	60	9.2
	10	80	11.8
	10	100	9.9
	30	10	24.4
	30	20	29.9
	30	30	30.1
	30	40	26.5
	30	60	26.2
	30	80	27.9
	30	100	27.6
	100	10	50
	100	20	60.9
	100	40	67.4
	100	60	81.5
	100	80	113.3
	100	100	92.7
	100	150	85.3

Note: We simulated 20 individuals from Pop A and 25 individuals from Pop B using MaCS coalescent simulator. The two populations diverged 1800 generations ago. The effective population size for both populations was set 12,500 at all times (except during the founder event). The mutation and recombination rates were set to $2x10^{-8}$ and $1x10^{-8}$ per base pair per generation. During the founder event, the effective population size reduced to 5 individuals for one generation at the date specified in the table above. For each simulation we generated data for ~450,000 polymorphic sites. SNPs with minor allele frequencies of <1% were discarded.

References

1. 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 Genetics 7, e1001373.

2. Hill, W., and Robertson, A. (1968). Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics 38, 226-231.

3. Chen, G.K., Marjoram, P., and Wall, J.D. (2009). Fast and flexible simulation of DNA sequence data. Genome Research 19, 136-142.

4. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Research 19, 318.

Appendix D

Supplementary Material for Chapter 5

Note D.1: Statistics used for estimating dates of admixture

Here we describe the *rolloff* and *ALDER* linkage disequilibrium (LD) statistics that we use for dating admixture events in India.

These related methods apply two key insights. The first is that mixture between two highly diverged populations creates LD that decays exponentially as recombination occurs: explicitly, as e^{-nd}, where n is the number of generations since admixture and d is the genetic distance between SNPs. The second insight is that the amount of admixture LD between each pair of SNPs is proportional to the product of the allele frequency differences between the ancestral populations.

The *rolloff* statistic introduced in Moorjani et al. (2011) estimates admixture LD by computing pairwise correlation between SNPs and weighting them by the differences in allele frequencies in reference populations[1; 2]:

$$A(d) = \frac{\sum_{|x-y| \approx d} z(x,y) w(x,y)}{\sqrt{\sum_{|x-y| \approx d} z(x,y)^2} \sqrt{\sum_{|x-y| \approx d} w(x,y)^2}} \qquad [6]$$

Here, *x, y* are SNPs separated by a distance *d* Morgans; *z(x,y)* = the correlation between SNPs *x* and *y*; and the weight function *w(x,y)* is the product of the allele frequency differences between the reference populations at *x* and *y*. We plot the weighted correlation with genetic distance and obtain a date by fitting an exponential function with a constant offset (affine) term: $y = Ae^{-nd} + c$, where n is the number of generations since admixture and d is the distance

in Morgans. Standard errors are computed using weighted block jackknife[3], with one chromosome dropped per run.

While this statistic provides accurate results under most scenarios, Moorjani et al. (2013) found that for groups that have a history of a very strong bottleneck after admixture, the normalization term $z(x,y)^2$ exhibits an exponential decay, thus biasing the estimated dates of admixture[4]. While this does not affect the estimates in outbred groups such as Europeans and Africans, it could cause a bias in the case of Indian groups such as Vysya and Chenchu that have a history of strong founder events in the past 100 generations.

Following the same correction as Moorjani et al. 2013, we apply two modifications to the *rolloff* statistic in equation [6]:

(a) $z(x,y)$ = the covariance between SNPs x and y. This makes the statistic more mathematically tractable, thus allowing us to use the amplitude of the exponential decay to estimate admixture proportions as in *ALDER*.

(b) Remove the normalization term $z(x,y)^2$ to remove the bias in the estimated dates of admixture.

The final statistic used is as shown in equation [2]. Simulations show that these changes provide accurate estimates of dates of admixture even in groups with a history of founder events[4].

The *rolloff* statistic requires access to estimates of the allele frequency differences between the reference groups to weight the SNPs and to make the statistic sensitive to admixture related LD. This means that we need data from reference groups that are related to the true ancestral populations. A challenge is that Ancestral South Indians (ASI) are not closely related to any extant group and are only anciently related to indigenous Andaman Islanders (Onge), who

provide poor estimates of ASI allele frequencies as their population size has been small for the tens of thousands of years since separation from the ASI. To overcome these limitations, we further modified the implementation of *rolloff*:

(a) *Use of SNP loadings estimated based on PCA as the weights in rolloff:* For India, we do not have access to samples of unadmixed ASI but we have access to multiple admixed groups differing in their mixture proportions. Thus, we can use SNP loadings from PCA of multiple admixed groups and a surrogate for ANI (say, Europeans) in place of the frequencies in equation [2]. This idea was first introduced in Moorjani et al. (2013), where simulations showed that PCA-based SNP loadings can be used to accurately infer dates[4].

(b) *Using the admixed group as one reference population:* Loh et al. (2013) recently extended ideas from *rolloff* in the *ALDER* method[5]. *ALDER* can infer admixture dates with just one reference population, and can also relate the amplitude of the fitted exponential to admixture proportions (see also Note D.4). Specifically, we compute the following statistic shown in equation [3]. Simulations show that this method provides accurate estimates for the date even when highly diverged ancestral groups are used as reference populations[5].

We applied both *rolloff* and *ALDER* to infer the dates of admixture using PCA-based loadings and single reference populations and show that we obtain qualitatively similar results from both methods (Table D.5).

<u>*Simulations using demographic parameters relevant to Indian groups*</u>*:*

Moorjani et al. (2011) reported that *rolloff* estimates can be upwardly biased in the cases of low admixture proportion and small sample sizes[1]. To evaluate how this might affect our

results in India, we created simulated chromosomes of mixed European and Asian ancestry for demographic parameters relevant to Indian groups. The choice of Europeans and East Asians as the ancestral populations for these simulations was motivated by the fact the $F_{st}(ANI, ASI)$ is approximately equivalent to $F_{st}(CEU, CHB) = 0.09$[6].

We simulated data based on the framework described in ref. [1]. For each Indian group (Brahmins, Mala, Pathan, *Dravidian rank 1* and *Indo-European rank 1*), we ran 100 simulations where we set the mixture proportion, time since mixture, and number of samples to match the parameters estimated for the specific group. Table D.6 shows that the estimated dates are within one standard error of the average of 100 simulations.

Note D.2: Test for multiple waves of admixture

Here we describe a method for identifying groups that have evidence for more than one wave of gene flow.

The method is based on a Likelihood Ratio Test (LRT) for whether the admixture LD decay curves fit the simple exponential decay expected for a single wave of admixture. For this purpose we use the output obtained from *rolloff* using PCA-based SNP loadings as the weights (Note D.1).

The *null* hypothesis is that there has been a single pulse of admixture. We use least squares to estimate the parameters of the null model by fitting $y = Ae^{-nd} + c$ where n = the date of admixture and d = genetic distance.

The *alternative* hypothesis is that there have been two pulses of admixture. We fit $y = Ae^{-n_1 d} + Be^{-n_2 d} + c$, where n_1 = date of the first pulse of admixture and n_2 = date of the second pulse of admixture. The log likelihood of each model is

$$\frac{-N}{2}(\log_e(2\pi) + 1 - \log_e(N) + \log_e(\sum_{i=1}^{n} \varepsilon_i^2)) \quad [7]$$

where, N = the number of data points in each simulation and ε_i = the residuals of the fitted model (true(y) - fitted(y)).

The difference between the log likelihood of the null vs. the alternative hypothesis (-2*$\log_e$(likelihood of null model) + 2*$\log_e$(likelihood of the alternate model)) is expected to be chi-square distributed with 2 degrees of freedom.

To test if the χ^2 approximation holds true in our case, we performed 100,000 numerical simulations of data under the null model of a single pulse of mixture (date range of 1-300 generations) with normal noise (mean = 0, standard deviation = 0.02, similar to the noise observed in real data). We then use least squares to estimate the parameters of the null ($y = Ae^{-nd} + c$) and alternative models ($y = Ae^{-n_1 d} + Be^{-n_2 d} + c$) and record the p-value of the likelihood ratio test assuming a χ^2 distribution with 2 d.f. We reject the null hypothesis in 5.7% of the simulations (Figure D.4).

We applied the LRT method to all groups with ≥10 samples (the requirement of a minimum sample size is motivated by the sensitivity of the test to noise in the case of few samples). For most Indo-European speaking groups, there is evidence to reject the null hypothesis (Table 5.2). In contrast, other groups can be reasonably well fit by the null model to within the limits of our resolution.

We conclude by highlighting three caveats of this LRT:

(1) Without comparing the model of two pulses of admixture with models of multiple pulses (>2) or gradual admixture, we cannot conclude that a group has a history of exactly two waves of admixture. In general, the true histories of the groups consistent with the null model almost certainly involved some amount of non-instantaneous gene flow, so with sufficiently high sample size, our test for a non-exponential decay would be almost guaranteed to reject the null model.

(2) A second caveat is that our method might reject the null due to sources of LD other than admixture such as LD due to founder events or ancestral LD. In theory, however, this problem is mitigated by using PCA loadings as weights.

(3) A third caveat is that autocorrelation across distant bins in *rolloff* will make our likelihood scores anti-conservative; we do not currently know how to correct for this autocorrelation. Thus, we treat the evidence of multiple-waves of admixture as suggestive only, and apply other formal methods to identify groups that are consistent with a single wave of ANI-ASI admixture.

Note D.3: Inferring the number of admixture events

Here we describe how we identified sets of Indian groups consistent with mixture of the same two ancestral populations within the limits of our resolution.

Our approach was first introduced in Reich et al. (2012) where it was applied to estimate the number of migrations from Siberia into the Americas[7]. Here, we co-analyze a panel of Indian groups (m) along with a panel of non-Indian groups (n). The idea is to compute f_4 statistics measuring the correlation in allele frequencies between each possible pair of Indian groups

(m(m-1)/2 comparisons), and each possible pair of non-Indian groups (n(n-1)/2 comparisons). Specifically, we compute statistics like $f_4(India_h, India_i; NonIndian_j, NonIndian_k)$. If the analyzed Indian groups harbor ancestry from exactly the same pair of ancestral populations ANI and ASI (but in different proportions), then the f_4 statistics should be proportional up to a scaling factor, and we can test this null hypothesis.

To implement this, we need to address the fact that many of the f_4 statistics can be written as linear combinations of each other, and thus we need to pick a basis for the space of f_4 statistics. In practice, we pick one Indian group as "$India_{base}$", and an African group (YRI) as "$NonIndian_{base}$" (the choice of base has no impact on the statistical findings). We then compute all possible f_4 statistics:

$$f_4(India_{base}, India_{other}; NonIndian_{base}, NonIndian_{other})$$

This yields a matrix of *m-1 x n-1* dimensions. Using Singular Value Decomposition (SVD), we estimate the number of independent components or *rank* of the expected values of the f_4 relationship matrix[7]. If the ANI and ASI ancestry in all the tested Indian groups derives from the same ancestral populations, the f_4 statistics measuring these correlations are expected to all be proportional, and thus the matrix will have one independent component or rank = 1. However, if a tested Indian group has a history of multiple gene flow events, the rank can be greater than 1, and we can test this null hypothesis using a Hotelling *T*-test. Our previously described methods also allow us to extend this to compute the minimum rank of the f_4 matrix needed to explain the data[7]. Assuming no back-migration from India into the panel of non-Indian groups, we can interpret a rank of r as implying at least $r+1$ ancestral populations.

Simulations

To test the method for demographic parameters relevant to Indian groups, we performed coalescent simulations using Hudson's *ms*[8]. For each simulation, we generated data for ~250K independent SNPs for 15 groups (Pop 1-15, 10 samples for each group). We set the effective population size (N_e) for all groups to be 12,500, and the mutation and recombination rates at $2x10^{-8}$ and $1x10^{-8}$ per base pair per generation respectively. Pop 1 is the outgroup that diverged from Pops 2 and 3 about 1,800 generations ago. Pop 2 and 3 diverged from each other 900 generations ago. The relationship of Pop 1, 2, and 3 can be considered analogous to the relationship of Yoruba, Onge, and CEU respectively. Pop 4-9 are related to Pop 3 analogously to the relationship of West Eurasians to ANI, and diverged from Pop 3 between 200 - 450 generations before the present.

Simulation 1: Single gene flow event with the same admixing populations

Consider the model in Figure D.5. Pops 10-15 are admixed and have ancestry from populations 2' and 3', which are closely related to Pop 2 and 3 respectively, with the admixed populations deriving between 20-80% ancestry from Pop 3. The date of admixture for all groups (Pop 10-15) is 100 generations before present. These groups are analogous to the groups on the "Indian cline" with the range of admixture proportions and dates set to be similar to those observed in real data (Figure D.5).

We estimate the rank of the f_4 relationship matrix f_4 (Pop 10-15; Pop 1-9). Here, Pop 10 is equivalent to $India_{base}$ and Pop 1 (an outgroup to Pop 2-15) is equivalent to $NonIndian_{base}$. We infer the number of independent components as 1 (rank 1 at P>0.05).

Simulation 2: Two gene flow events involving different ancestral populations

Pop 10-15 are admixed and Pop 10-14 have ancestry from populations 2' and 3' with Pop 3' ancestry varying between 20-80%. However, Pop 15 has 35%/65% ancestry from Pop 4 / Pop 2. All admixture events occurred 100 generations ago. We estimate the rank of the f_4 relationship matrix as f_4(Pop 10-15; Pop 1-9) and infer number of independent components is 2. However, if we remove Pop 15 from the analysis, that is, f_4(Pop 10-14; Pop 1-9), the inferred rank is 1, as expected.

Simulation 3: Three independent gene flows with different mixing populations

Pop 10-15 are admixed and Pop 10-13 have ancestry from populations 2' and 3', with Pop 3' ancestry varying between 20-80%. However, Pop 14 has 70%/30% Pop 5 / Pop 2 ancestry, and Pop 15 has 35%/65% Pop 4 / Pop 2 ancestry. All admixture events occurred 100 generations ago. We estimate the rank of the f_4 relationship matrix f_4(Pop10-15; Pop1-9) and infer the number of independent components is 3.

Simulation 4: Two independent gene flow events at different time periods

Pop 10-15 are admixed and have ancestry from populations 2' and 3', with Pop 3' ancestry varying between 20-80%. Admixture occurred 100 generations ago. However, Pop 15 also has ancestry from an older gene flow event that occurred 150 generations ago with 50% Pop2' / 50% Pop3' ancestry. Thus overall, Pop 15 has 70%/30% ancestry from Pop 2' / Pop 3'. We estimate the rank of the f_4 relationship matrix f_4(Pop 10-15; Pop 1-9) and infer the number of independent components is 2.

Simulation 5: Multiple independent gene flow events at different time periods

Pop 10-15 are admixed and have ancestry from populations 2' and 3', with Pop 3' ancestry varying between 20-80%. Admixture occurred 50-300 (intervals of 50) generations ago (such that Pop 10 was admixed 50 generations ago, Pop 11 was admixed 100 generations ago, etc.). The f_4 relationship matrix is f_4(Pop 10-15; Pop 1-9) and we infer the number of independent components is 3.

In conclusion, our simulations demonstrate that we can accurately estimate the minimum number of gene flow events, and post-admixture drift alone does not change the rank of the f_4 relationship matrix.

Results

We performed a systematic analysis to identify groups that have a similar history of ANI-ASI mixture, meaning that all their ancestry is consistent with being derived from exactly the same ANI and ASI ancestral populations to within the limits of our resolution. We restrict this analysis to all Indian groups that have at least 5 samples and all non-Indian groups that have at least 10 samples, including groups from East Asia, Europe, the Middle East, the Caucasus, and Africa. We remove all Central Asian and South Asian populations from the list of Non-Indian groups as these have an increased likelihood of back-migration from India in the recent past that can complicate interpretation. We include Vedda (4 samples), an aboriginal population from Sri Lanka, as this population appeared to have a relatively simple history of ANI-ASI mixture in a preliminary analysis. The analyzed data thus consists of m=37 Indian groups (including Onge) and n=38 non-Indian groups.

To identify sets of Indian populations that are consistent with deriving all their ancestry from exactly the same ANI-ASI ancestral populations, we systematically explored sets of these Indian groups. We used an iterative procedure, as follows:

(1) Testing all possible sets of three Indian groups

We start by computing:

$$f_4(India_{set\ of\ three\ groups};\ Yoruba,\ NonIndian_{other})$$

and estimate the ranks of the resulting 2 x *n-1* matrix using a likelihood ratio test. We repeat this for all (37)(36)(35)/6 possible triples of Indian groups.

For each triple of Indian groups that is consistent with a simple mixture of ANI and ASI (rank 1 at $P > 0.05$), we performed a further level of stringent testing for whether the proposed model is consistent with our data. Specifically, for triples of Indian groups that are consistent with being rank 1, we also run the *admixture graph* phylogeny-testing software[2; 6] to test if the set of population relationships shown in Figure D.1 with Pop 1 = Georgians and Pop 2 = Basque are consistent with the data to within the limits of our resolution (this is the same set of groups we use for estimating ancestry proportions in *F_4 Ratio Estimation* and thus here we are formally testing whether the model of history underlying the *F_4 Ratio Estimates* is valid). To evaluate significance, we use the criterion than none of the *f_2, f_3 and f_4*-statistics relating the 5 analyzed groups in the admixture graph is more than three standard errors from expectation.

(2) Testing sets of 4 Indian groups

For all triples of India groups that pass these two tests, we advanced to the next round, testing sets of four Indian groups for consistency with being a simple mixture of exactly the same ANI and ASI ancestral populations. Specifically, we took each of the passing triples and added in turn each of the remaining groups that were part of at least one triple that was rank 1. We applied the same two tests for consistency with a simple ANI-ASI mixture, leading to passing quadruples.

(3) Testing sets of 5, 6, and 7 Indian groups

We applied the same procedure to test higher numbers of groups. The results of each round are recorded in Table D.7 and Table D.8. We stopped finding sets of groups that pass the test after m=6.

We highlight two qualitative results that emerge from this analysis:

- Onge is often included among the sets of groups that are consistent with being rank 1, suggesting that it is consistent with being an ancient sister group for ASI as we previously suggested ref. [9]. However, for some sets of Indian groups qualifying as rank 1, we cannot add in Onge, suggesting that there also might be differences in ASI ancestry within India.
- A higher proportion of sets including lower caste and tribal populations have rank 1 than sets including upper caste groups.

Note D.4: Test for a single wave of admixture: comparison of predicted and observed *ALDER* amplitudes

To evaluate whether the admixture LD we are detecting in India could plausibly reflect a single wave of gene flow accounting for all the ANI-ASI mixture, we compared the observed amplitude of LD decay and the *ALDER* theoretical expectation for a model of single wave of mixture[5].

We run *ALDER* with one reference population (X or Y) and plot the weighted covariance against genetic distance and perform a least-squares fit using $y = Ae^{-nd} + c$, where n is the number of generations since admixture and d the genetic distance in Morgans. Under a single-wave mixture model, the amplitude of admixture LD decay, defined as $a_o = A + c/2$, is analytically predicted by the ANI ancestry proportion (α) using the relationship shown in equation [4] (See population relationships in Figure D.2).

The genetic drift ($f_2(ANI,X")$ and $f_2(ASI,X")$) (Figure D.2) can be obtained by fitting a model of population relationships using *admixture graph* to the data for an analyzed set of populations. By comparing the observed amplitude inferred from LD (measured with *ALDER*) and the expected amplitude from frequency correlations (using *admixture graph* or *F_4 Ratio Estimation*, which use similar information), we can infer how much of the total ANI ancestry in each Indian group is due to mixture in the last few thousand years.

We applied this analysis to two sets of Indian groups, an *Indo-European rank 1* set consisting of 4 groups and a *Dravidian rank 1* set consisting of 5 groups. We chose these from all the sets identified in Note D.3 based on two criteria:

(a) All groups are genotyped on the Affymetrix 6.0 array. This allows us to use more SNPs (n=210,482 SNPs) thus improving the accuracy of *ALDER*, and to include Onge, an essential population for our *admixture graph* analysis.

(b) The groups in the sets span as large a range as possible of ANI ancestry, which is valuable for constraining internal branch lengths in *admixture graph*.

Based on these criteria, we chose the following two sets:

Indo-European rank 1 set (n=4 groups; 32 samples):

Bhil, Jain, Lodi, Tharu

Dravidian rank 1 set (n=5 groups, 33 samples):

Adi-Dravidar, Kuruchiyan, Madiga, Malai Kuravar, Narikkuravar

We used the population relationships shown in Figure D.1, but now with only one West Eurasian outgroup (because we do not have access to Georgians on the Affymetrix array), as input to *admixture graph*. We confirmed that the Indo-European (n=32) and Dravidian (n=33) rank 1 sets are still good fits to the proposed model using the larger number of SNPs (n=210,482 rather than n=86,213 used in Note D.3). Specifically, none of the f_2, f_3 and f_4-statistics comparing all possible sets of groups are more than three standard errors from the model-based expectation.

The fit generated by *admixture graph* allows us to estimate the genetic drift that has occurred between: (1) ANI and the population X" that was ancestral to ANI and the sister group (X) we use in our *admixture graph* analysis (we tried a range of West Eurasian groups X), and (2) ASI and the population that was ancestral to ASI and the sister group we use for them (Onge)

(Figure D.2). We are able to estimate these branch lengths as we have access to several admixed populations that we hypothesize descend from the same admixture event.

We apply this method to compare the predicted amplitude of the admixture LD from *admixture graph* (based on allele frequency correlation information and the expectation shown in equation [4]) to the observations from *ALDER*, for a variety of proposed West Eurasian outgroups X, and for the Indo-European speaking set of rank 1 and Dravidian rank 1 groups (Table 5.3).

A complication of having only a single West Eurasian outgroup in the *admixture graph* is that it causes the model to be poorly constrained, but we can compensate for this by fixing the value of the admixture proportion (α) to be equal to the ANI ancestry inferred from *F_4 Ratio Estimation* using the merged Illumina-Affymetrix dataset. In this merged dataset, we have access to two West Eurasian outgroups which allows us to obtain precise ancestry estimates. We use Georgians and Basque, based on the *admixture graph* testing of Note D.3, and observe that this model provides a good fit to the data for many Indian groups.

To test if the expected amplitude based on the model of single admixture is consistent with the observed amplitude of admixture LD, for a tested set of Indian groups we measure the difference between two quantities:

Expected amplitude: We use *F_4 Ratio Estimation* on the set of Indian groups to obtain a point estimate of the admixture proportion, and we use *admixture graph* analysis on the same set of Indian groups (using the constrained model described above) to infer the genetic drift lengths $f_2(ANI,X")$ and $f_2(ASI,X")$. Plugging these numbers into the *ALDER* amplitude formula (equation [4]) provides a precise mathematical expectation for the

amplitude of admixture LD for the scenario that all the ANI-ASI admixture is due to a single admixture event.

Observed amplitude: We obtain this by performing *ALDER* analysis for the same pool of Indian groups using Basque as the reference population.

(Observed - Expected) amplitude: We compute the difference in amplitudes.

To infer statistical uncertainty, we use a weighted block jackknife dropping each chromosome in turn and repeating the entire procedure. This produces a standard error and allows us to test whether the difference between the expected and observed amplitude is consistent with zero (consistent with the null model of single wave of ANI-ASI admixture). Since the difference is consistent with zero and negative values are not genetically relevant, we can also compute a one-sided 95% confidence interval of 0% to mean + 1.65 times the standard error.

In practice, we did not find significant evidence for a difference between the observed and expected amplitudes in India. However, it is also interesting to place an upper bound on the proportion of ANI ancestry that could possibly derive from an earlier wave of admixture. To do this, we consider the alternative hypothesis that there were two waves of admixture and infer the maximum proportion of ANI ancestry that could possibly be unexplained by the dated ANI-ASI admixture.

Specifically, the model we are considering here is two waves of admixture from ANI-related ancestral populations that can be assumed to have the same allele frequencies, such that the older wave of mixture is old enough that its contribution to the measured LD is negligible. In this model, present-day Indian groups derive their ancestry from three sources: old ANI (α_{old}),

recent ANI (α_{new}), and ASI ($1-\alpha_{total}$). Hence, the second wave of ANI ancestry (proportion α_{new}) enters an admixed population (proportion $1-\alpha_{new}$) whose allele frequencies can be written as a linear combination from the first wave:

$$\alpha_{old}/(1-\alpha_{new}) * A + (1-\alpha_{old}/(1-\alpha_{new})) * B \qquad [8]$$

The expected one-reference-population *ALDER* amplitude shown in equation [4] then becomes:

$$a_o = \frac{2\alpha_{new}(1-\alpha_{new})(1-\alpha_{total})^2}{(1-\alpha_{new})^2}(\alpha_{total} f_2(ANI,X") - (1-\alpha_{total}) f_2(ASI,X"))^2 \qquad [9]$$

This reduces to the form shown in equation [5] shown earlier.

$$a_0 = \frac{2\alpha_{new}(1-\alpha_{total})^2}{\alpha_{old} + (1-\alpha_{total})}(\alpha_{total} f_2(ANI,X") - (1-\alpha_{total}) f_2(ASI,X"))^2 \qquad [10]$$

The last squared factor remains the same as in the single-wave case because we have assumed that the two ANI populations have the same allele frequencies. Note that replacing $\alpha_{old} = 0$ (so that $\alpha_{new} = \alpha_{total}$) reduces equation [5] to equation [4]. This amplitude is lower than the corresponding value for a single wave of admixture, since the admixture LD due to the older wave is no longer detectable beyond the shortest genetic distances. Thus, if the observed amplitude is lower than the expected (single-wave) amplitude, we can find the value of α_{old} that would explain the difference under a two-wave model.

Simulations

We now demonstrate the accuracy of the above line of reasoning in scenarios simulated to resemble hypothetical admixture histories of India. To capture some of the complexities of real human populations, we built our simulated data sets using phased data from the real groups from HGDP and HapMap using the method described in ref. [1]. Specifically, we simulated two sets of admixed groups as follows:

- Set 1: Three groups with [30%, 50%, 70%] ancestry from Europe (HapMap CEU) and the remaining ancestry from East Asia (HGDP Han).
- Set 2: Three groups with [20%, 30%, 40%] ancestry from Europe and the remaining ancestry from East Asia.

For each group, we generated 14 diploid individuals under two alternative admixture histories:

- A single CEU–Han admixture event 100 generations ago.
- Two waves of CEU admixture into Han, 300 and 75 generations ago, that together produce the same total fraction of CEU ancestry as shown above.

To perform the *admixture graph* analysis, we require additional outgroups. For this we use real data from HGDP French, Basque, Yoruba, and Dai and use the model shown in Figure D.6. We use the drift lengths and admixture proportions estimated by *admixture graph* to compute the expected amplitude (Note: The constrain of fixing the admixture proportion from F_4 *Ratio Estimation* is not required as we have two West Eurasian outgroups here). We performed *ALDER* single-reference analysis for each set of admixed groups with Basque and Dai reference populations (independently). We note that we do not reuse the CEU or Han populations (which were used to generate the simulated data) in our inference procedure, to account for the fact that

we do not have access to the true ancestral populations (ANI and ASI) for India. We created simulated populations in groups of three to allow us to infer the necessary f_2 values in the amplitude formula, as these are otherwise confounded.

We designed our simulations to match the parameters observed in India. In Figure D.6, the four outgroups (French, Basque, Yoruba, and Dai) take the place of Georgians, Basque, Yoruba, and Onge in Figure D.1. For the simulated histories, we chose a date of 100 generations before the present to be similar to the observed average age of ANI–ASI admixture in India. The two-wave dates of 300 and 75 generations ago provides a plausible alternative scenario yielding a *ALDER* curve similar to that of 100 generations. Finally, the two simulated population sets covered distinct ranges of the admixture proportion space, one with larger CEU ancestry components and higher ancestry proportion variation than the other. For both the sets, our inference methods provide reliable results.

Our simulation results demonstrate that for a single-wave admixture history, the weighted LD amplitude measured by *ALDER* is consistent with the expectation of our formula, whereas in the case of two-wave admixture, the measured *ALDER* amplitude is smaller than the expectation, as claimed in equation [4] (Table D.9). Out of the 12 population-reference pairs, the difference in the amplitudes is statistically consistent with zero ($|Z|<3$) for all single-wave simulations, while the difference is significantly different from zero in 8 of the 12 two-wave simulations, including all 6 with Basque as the reference population. Correspondingly, the estimates of the mixture proportion α_{old} required to explain the amplitude discrepancy under the alternate model are considerably smaller for the single-wave simulated data, with zero always included in the confidence interval (Table D.9).

Results

(a) *Indo-European rank 1* groups

For all West Eurasian groups, the model of population relationships provides a good fit to the Indo-European rank 1 data as assessed by *admixture graph* (such that none of the *f*-statistics are greater than three standard errors from expectation). Thus we plug in the admixture proportions and the drift lengths ($f_2(ANI,X")$ and $f_2(ASI,X")$) computed using *admixture graph* in equation [4] to estimate the expected amplitude. We observe that the expected amplitude is consistent with the observed amplitude in *ALDER* ($|Z| < 3$ for a difference between the two estimates over all 7 West Eurasian groups we tested) (Table 5.3).

For subsequent analyses we focused on Basque as the reference population and fixed the ANI ancestry proportion from *F_4 Ratio Estimation* as described above. We compute the difference in amplitude (observed - expected) and find that the two estimates are statistically consistent (Z = -0.35), suggesting that the model of single wave of ANI-ASI admixture is consistent with our data.

Applying the alternate two-wave amplitude formula (equation [5]), we estimate the range of possible α_{old} as 4.5 ± 8.5%, with a 95% confidence interval of 0-18.6% (truncated at 0). Thus we find no evidence to reject a single-wave model with all ancestry contributing to the measured admixture LD.

(b) *Dravidian rank 1* groups

Similar to the *Indo-Europeans rank 1* groups, we applied *admixture graph* and *ALDER* to the *Dravidian rank 1* data using various West Eurasian groups as references and found that the

expected and observed amplitudes are consistent ($|Z| < 3$ for all reference populations tested) (Table 5.3).

We focused next on Basque as the reference population and used the ANI ancestry proportion estimated from *F_4 Ratio Estimation.* We observe that the expected amplitude is consistent with the observed amplitude in *ALDER* (Z= -1.06), suggesting that the model of single wave of ANI-ASI ancestry provides a fit to the data. The proportion of ANI ancestry unexplained by our model (α_{old}) is 7.1 ± 5.5%, with a 95% confidence interval of 0-16.2% (truncated at 0).

In conclusion, our data are consistent with the null model of a single wave of ANI-ASI admixture in the history of selected Indo-European and Dravidian speaking groups in India.

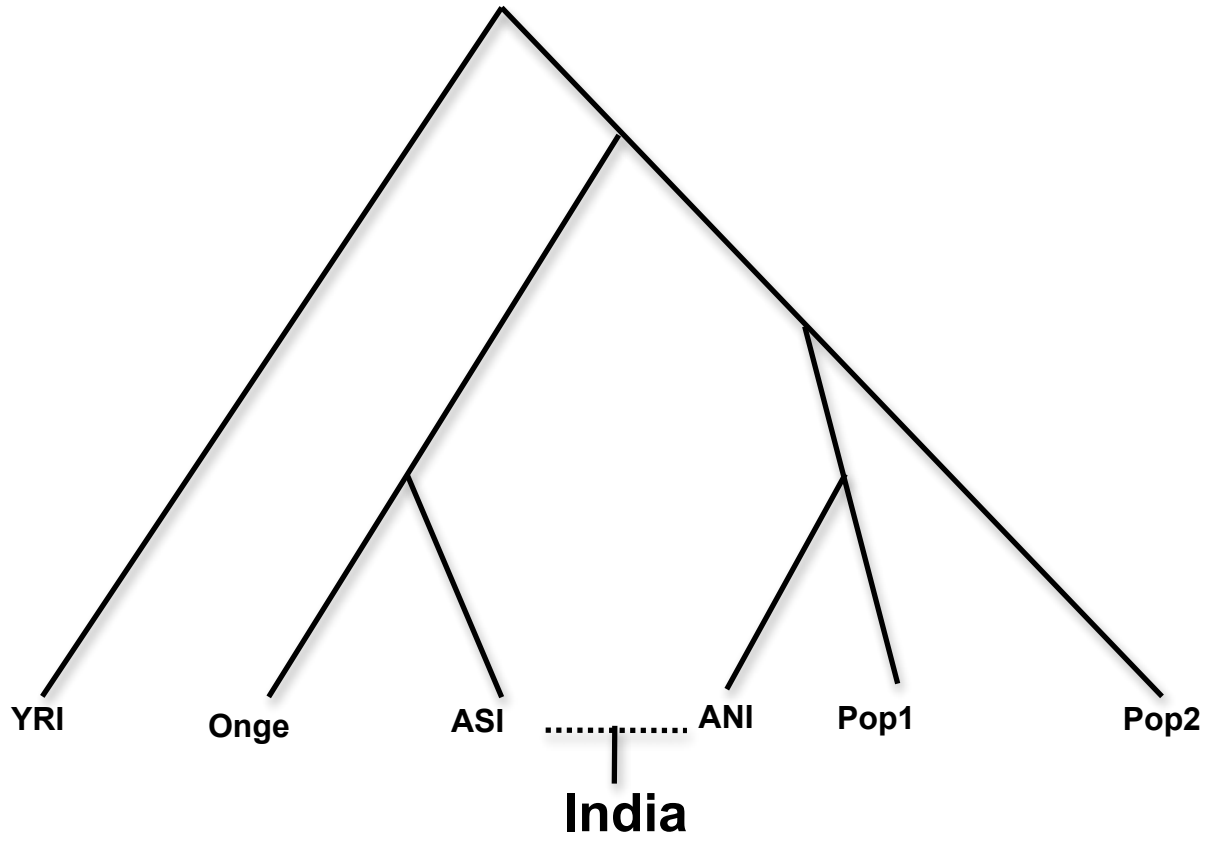
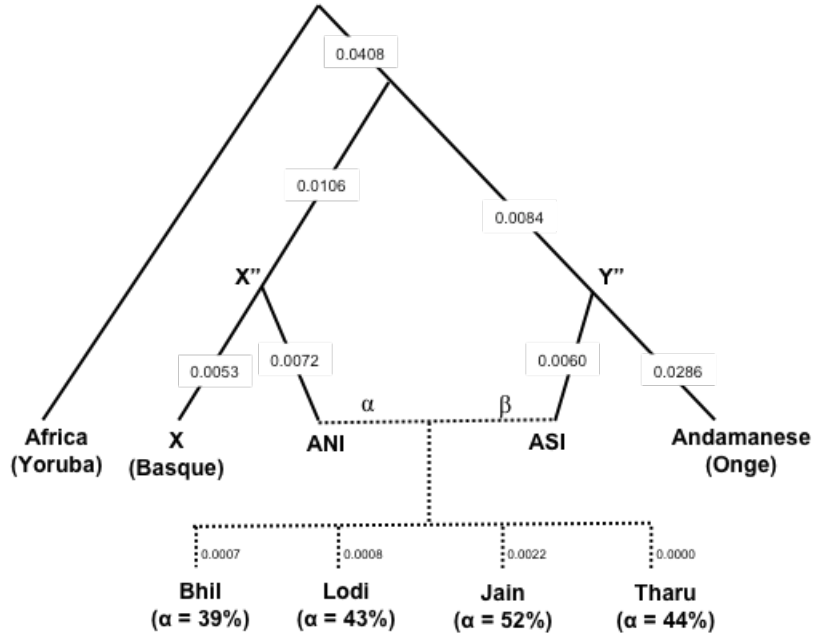


Figure D.1. Historical relationships assumed for *F_4 Ratio Estimation*

(a) Indo-European rank 1 set



(b) Dravidian rank 1 set

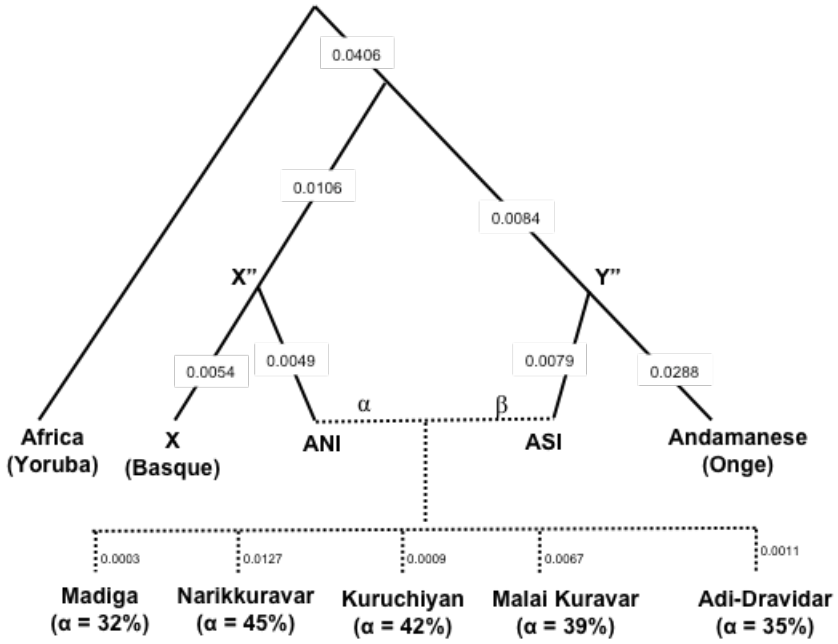


Figure D.2. ***admixture graph*** **fitted models of Indian history.**

(a) Full dataset

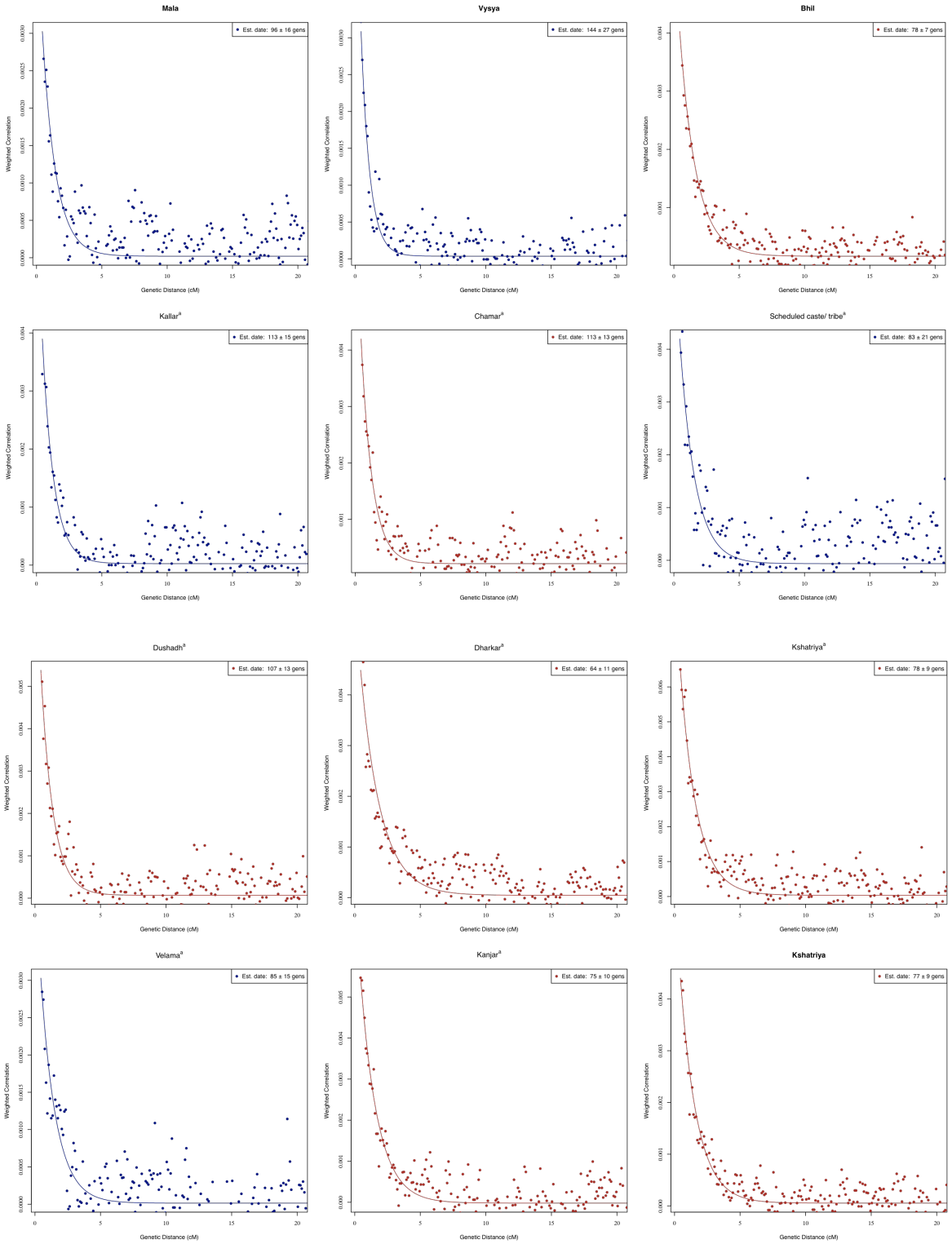
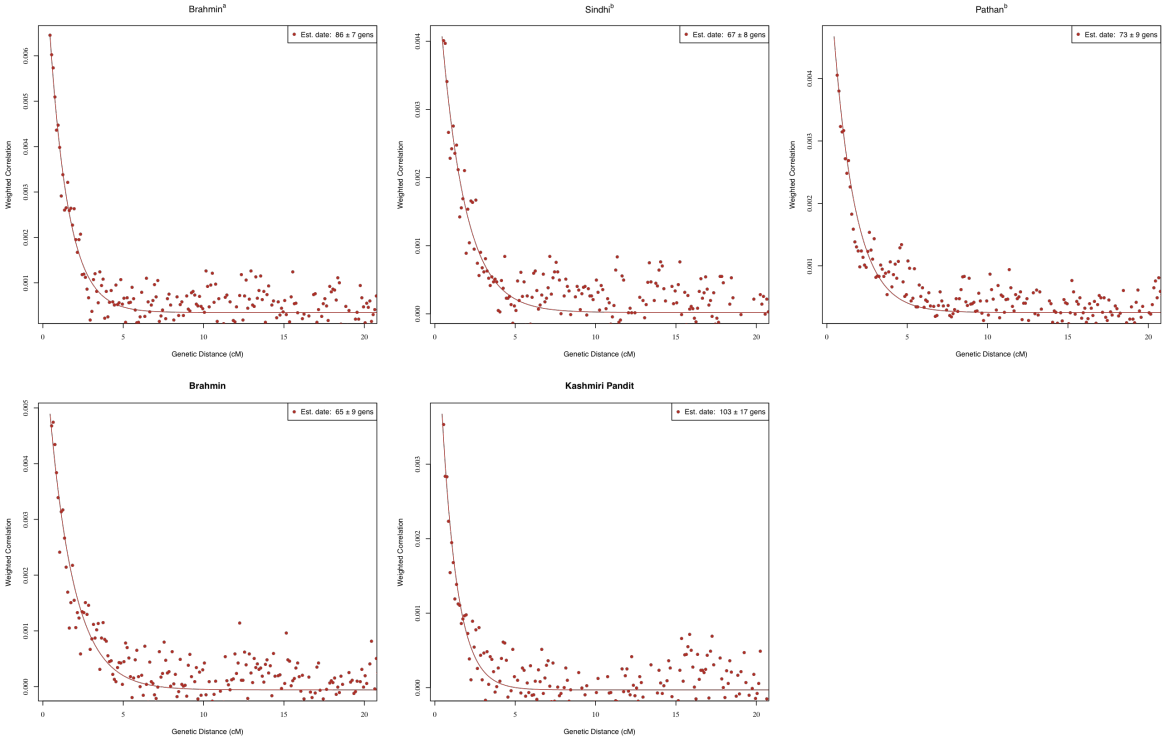


Figure D.3. *rolloff* curves specific to each Indian group.

Figure D.3. (Continued)



(b) *Rank 1* groups with a simple history of ANI-ASI admixture

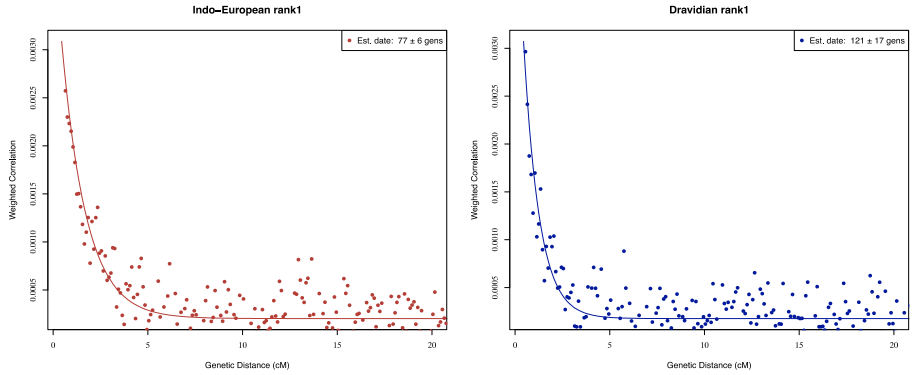


Figure D.3. (a) We use the full Affymetrix (494,863 SNPs) or Illumina (500,703 SNPs) dataset to increase precision. We run *rolloff* with weights computed by performing Principal Component Analysis (PCA) on data from all populations on the Indian-cline and CEU (excluding the test population). Samples are colored by linguistic affiliation. ***(b)*** We also performed *rolloff* analysis for *rank 1* groups, computing PCA based SNP loadings for Basque and Indian cline groups (not including the target admixed groups for computing the weights) with data for 210,482 SNPs. We ignore inter-SNP distances less than 0.5 cM to avoid confounding by background LD.

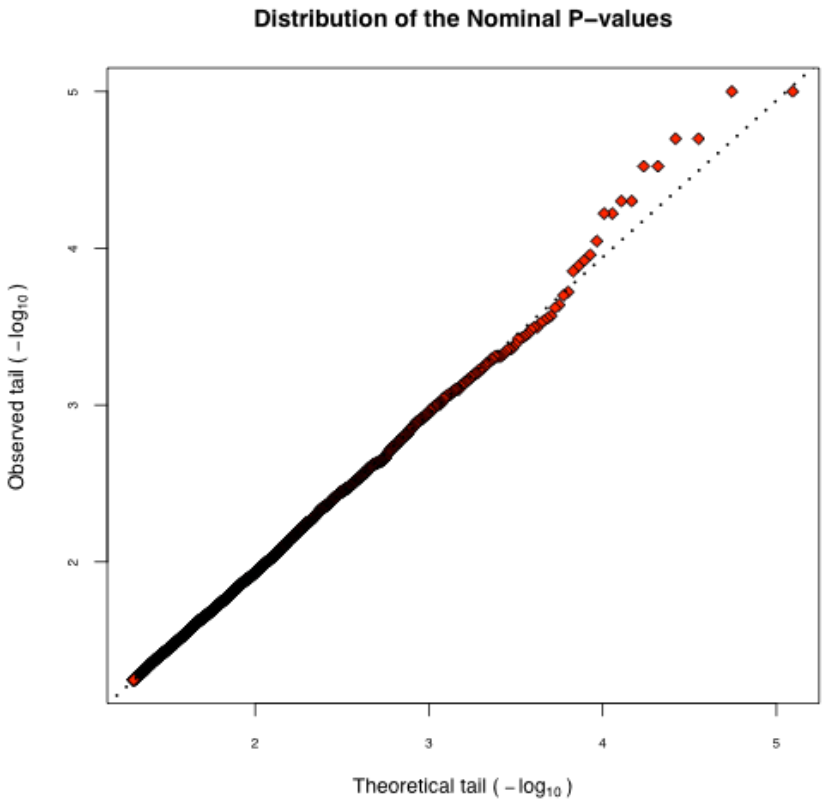


Figure D.4. Distribution of nominal p-values in simulations for likelihood ratio test. We performed 100,000 simulations based on the null model of a single pulse of mixture with noise. We use least squares to fit a null model of a single pulse of mixture ($y = Ae^{-nd} + c$) and an alternative model of two pulses of admixture ($y = Ae^{-n_1 d} + Be^{-n_2 d} + c$), where n, n_1, n_2 are parameters capturing the times since mixture, and d is the genetic distance. We performed a likelihood ratio test with 2 degrees of freedom and plotted the distribution of the nominal p-values. We computed the observed tail (y) as the proportion of observed p-values that are less than or equal to the theoretical p-values (x), normalized by the total number of simulations. Values below $-\log_{10}(0.05)$ are not shown. The dotted line indicates the regression line for the linear model between $\log_{10}(y)$ and $\log_{10}(x)$.

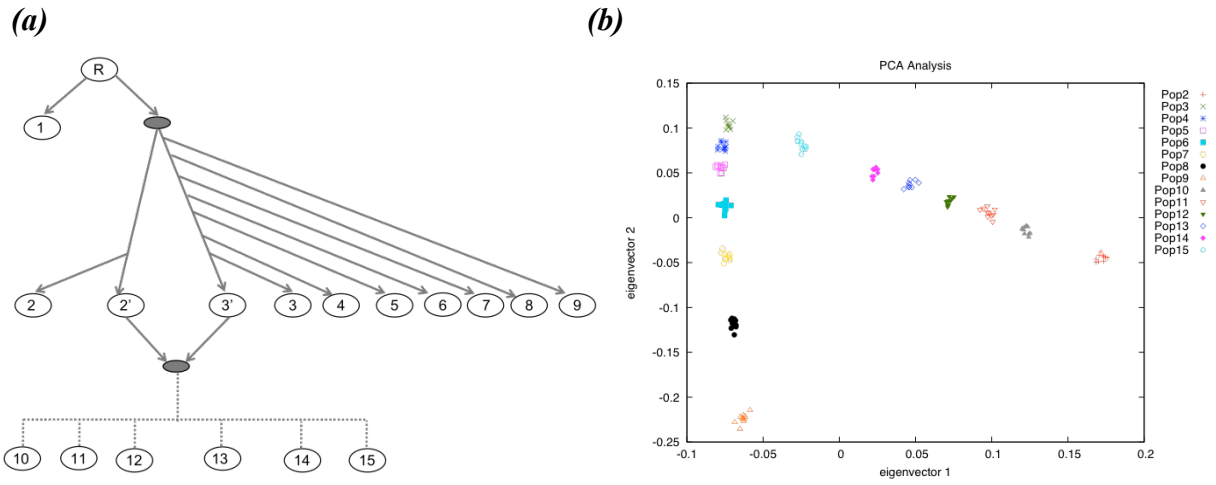


Figure D.5 Phylogenetic relationships of simulated populations Pops 1-15. Panel (a) shows the phylogenetic relationships of Pops 1-15, and (b) shows PCA of Pops 2-15. SNPs were ascertained in Pop 1 and hence this population is not included in the PCA.

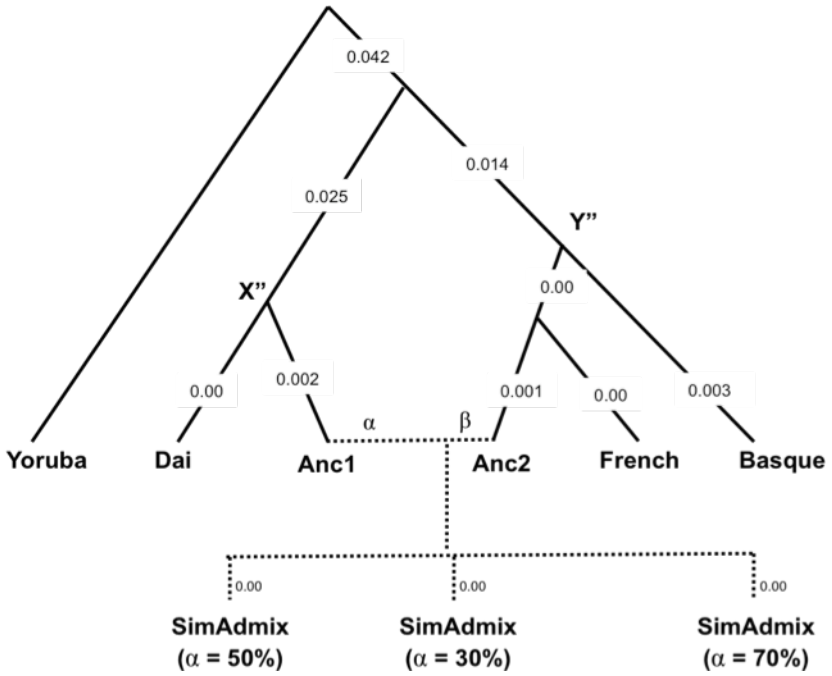


Figure D.6. ***Admixture graph*** **for simulated data used in** ***ALDER***

Table D.1: Data curation

Pop	Dataset	Samples removed pre-PCA curation	Samples removed post-PCA curation	Total count (Post curation)	Exclusion Criterial[c]
Adi-Dravidar	This study			5	
Bhil	Reich 2009 & this			17	
Bhumij	This study		5	0	(1)
Birhor	This study	1	4	0	(a); (1)
Brahmin	This study	5		10	(b)
Changpa	This study		5	0	(1)
Gond	This study	1	14	0	(a); (1)
Ho	This study		5	0	(1)
Irula	This study		5	0	(2)
Jain	This study			5	
Jews	This study		5	0	(2)
Kallar	This study			5	
Kattunayakan	This study			5	
Korku	This study	1	4	0	(a); (1)
Kshatriya	This study	5		15	(b)
Kuruchiyan	This study			5	
Gounder	This study			5	
Madiga	Reich 2009 & this	5	1	13	(b); (3)
Malai Kuravar	This study			5	
Mala	Reich 2009 & this	5		13	(b)
Mali	This study			5	
Minicoy	This study		1	4	(3)
Munda	This study		5	0	(1)
Narikkuravar	This study			5	
Palliyar	This study			5	
Kashmiri Pandit	Reich 2009 & this	5		15	(b)
Paniya	This study			5	
Sherpa	This study		5	0	(1)
Siddi	Reich 2009 & this	2	14	0	(a); (1)[6; 10]
Subba	This study		5	0	(1)
Tibet-refugees	This study		5	0	(1)
Vedda	This study			4	
Vysya	Reich 2009 & this	5	1	14	(b); (3)
Tharu	Reich 2009		4	5	(3)
Meghawal	Reich 2009			5	
Chenchu	Reich 2009			6	
Kurumba	Reich 2009		9	0	(2)
Hallaki	Reich 2009		7	0	(2)
Santhal	Reich 2009		7	0	(1)
Kharia	Reich 2009		6	0	(1)
Vaish	Reich 2009			4	
Srivastava	Reich 2009			2	
Naidu	Reich 2009			4	
Velama	Reich 2009			4	
Sahariya	Reich 2009		4	0	(1)
Lodi	Reich 2009			5	
Satnami	Reich 2009		1	3	(3)
Kamsali	Reich 2009			4	
Onge	Reich 2009			9	
Great_Andamanese	Reich 2009		7	0	(1)[6]
Nyshi	Reich 2009		4	0	(1)
Ao Naga	Reich 2009		4	0	(1)
Brahmin[a]	Metspalu 2011			8	
Kanjar[a]	Metspalu 2011	1		8	(b)
Chamar[a]	Metspalu 2011			10	

Table D.1 (Continued)

Pop	Dataset	Samples removed pre-PCA curation	Samples removed post-PCA curation	Total count (Post curation)	Exclusion Criterial[c]
Dushadh[a]	Metspalu 2011	3		7	(b)
Kshatriya[a]	Metspalu 2011			7	
Kol[a]	Metspalu 2011		16	0	(1)
Dharkar[a]	Metspalu 2011	1		11	(b)
Muslim[a]	Metspalu 2011			5	
Scheduled caste[a]	Metspalu 2011		6	0	(2)
central_mix1_nihali	Metspalu 2011		5	0	(2)
Gond[a]	Metspalu 2011		4	0	(2)
Munda[a]	Metspalu 2011		1	0	(1)
Scheduled caste/ tribe[a]	Metspalu 2011			6	
Hakkipikki[a]	Metspalu 2011			4	
Ao Naga[a]	Metspalu 2011		4	0	(1)
Chenchu[a]	Metspalu 2011			4	
Kallar[a]	Metspalu 2011			8	
Velama[a]	Metspalu 2011	1		9	(a)
Palliyar[a]	Metspalu 2011			5	
Sindhi[b]	Li 2008		14	10	(1)[6]
Pathan[b]	Li 2008		7	15	(1)[6]

[c]Samples were removed based on the following exclusion criteria:
Pre-PCA curation:
(a) Remove all duplicate samples based on >90% matching with another sample in the dataset.
(b) Remove all related samples: In case of trios, child was excluded and in case of first-degree relative, one sample from the pair was excluded.
(c) Removed all samples previously excluded in Metspalu et al (2011): 8 samples excluded (not shown in table above)[11].

Post-PCA curation (Figure 5.1):
(1) Remove samples and groups that have evidence of recent ancestry from groups other than ANI and ASI based on PCA.
(2) Remove groups that are not homogenous in PCA.
(3) Remove samples that do not cluster with the majority of samples from their group.

Table D.2: Summary of D-statistics

Population (X)	*n*	Language Group	Social/ caste group	Pop with highest D-statistic mean	Pop with 2nd Highest D-statistic	Pop with 3nd Highest D-statistic
Paniya	5	Dravidian	Tribal	Georgian	Armenian (0.1)	Cypriot (0.3)
Palliyar	5	Dravidian	Tribal	Georgian	Kurd (0.1)	Abhkasian (0.2)
Kattunayakan	5	Dravidian	Tribal	Georgian	Kurd (0.2)	Armenian (0.4)
Palliyar[a]	5	Dravidian	Lower caste	Cypriot	Abhkasian (0.4)	Georgian (0.5)
Madiga	13	Dravidian	Lower caste	Georgian	Lezgin (1.1)	Abhkasian (1.2)
Mala	13	Dravidian	Lower caste	Georgian	Abhkasian (0.5)	Armenian (1.2)
Adi-Dravidar	5	Dravidian	Lower caste	Georgian	Abhkasian (1.1)	Armenian (1.5)
Hakkipikki[a]	4	Dravidian	Tribal	Georgian	Armenian (1.2)	Abhkasian (1.3)
Vedda	4	Indo-European	Tribal	Georgian	Abhkasian (0.1)	Kurd (0.4)
Kamsali	4	Dravidian	Lower caste	Georgian	Armenian (0.8)	Lezgin (0.6)
Chenchu[a]	4	Dravidian	Tribal	Georgian	Abhkasian (1.4)	Armenian (2)
Chamar[a]	10	Indo-European	Tribal	Georgian	Abhkasian (0.6)	Lezgin (1.1)
Chenchu	6	Dravidian	Tribal	Georgian	Abhkasian (0.7)	Armenian (1.2)
Bhil	17	Indo-European	Tribal	Georgian	Abhkasian (0.4)	Armenian (1)
Kallar	5	Dravidian	Lower caste	Georgian	Armenian (0.9)	Cypriot (1)
Kallar[a]	8	Dravidian	Tribal	Georgian	Abhkasian (1.2)	Lezgin (2)
Vysya	14	Dravidian	Middle caste	Georgian	Abhkasian (1.1)	Armenian (1.7)
Malai Kuravar	5	Dravidian	Tribal	Georgian	Abhkasian (0.4)	Armenian (1.5)
Satnami	3	Indo-European	Lower caste	Georgian	Abhkasian (0.2)	Tuscan (0.8)
Kuruchiyan	5	Dravidian	Tribal	Georgian	Abhkasian (1.1)	Armenian (1.7)
Dushadh[a]	7	Indo-European	Lower caste	Georgian	Abhkasian (0.2)	Lezgin (1.1)
Sch.caste/ tribe[a]	6	Dravidian	Lower caste	Georgian	Kurd (0.6)	Abhkasian (0.8)
Mali	5	Dravidian	Lower caste	Georgian	Armenian (1.3)	Abhkasian (1.4)
Minicoy	4	Indo-European	Lower caste	Georgian	Abhkasian (2.3)	Cypriot (2.3)
Gounder	5	Dravidian	Middle caste	Georgian	Kurd (0.6)	Abhkasian (0.9)
Lodi	5	Indo-European	Lower caste	Georgian	Armenian (2)	Abhkasian (2)
Naidu	4	Dravidian	Upper caste	Georgian	Armenian (1.3)	Abhkasian (1.1)
Velama	4	Dravidian	Upper caste	Georgian	Abhkasian (1.3)	Armenian (2.8)
Velama[a]	9	Dravidian	Upper caste	Georgian	Armenian (1)	Abhkasian (0.9)
Narikkuravar	5	Dravidian	Tribal	Georgian	Abhkasian (0.5)	Cypriot (1.3)
Tharu	5	Indo-European	Tribal	Georgian	Lezgin (0.8)	Abhkasian (1.1)
Dharkar[a]	11	Indo-European	Nomadic group	Georgian	Tuscan (1)	Armenian (2)
Kanjar[a]	8	Indo-European	Nomadic group	Georgian	Abhkasian (1.2)	Tuscan (0.8)
Muslim[a]	5	Indo-European	Religious group	Georgian	Abhkasian (1.2)	Armenian (1.9)
Srivastava	2	Indo-European	Upper caste	Georgian	Abhkasian (-0.3)	Cypriot (0.1)
Jain	5	Indo-European	Religious group	Georgian	Abhkasian (0.8)	Lezgin (1.2)
Meghawal	5	Indo-European	Lower caste	Georgian	Abhkasian (0.8)	Cypriot (1.3)
Kshatriya[a]	7	Indo-European	Upper caste	Georgian	Abhkasian (1.4)	Lezgin (1.9)
Vaish	4	Indo-European	Upper caste	Georgian	Lezgin (1)	Armenian (2.2)
Brahmin[a]	8	Indo-European	Upper caste	Tuscan	Lezgin (0)	Georgian (0.1)
Kshatriya	15	Indo-European	Upper caste	Georgian	Abhkasian (0.9)	Tuscan (0.8)
Brahmin	10	Indo-European	Upper caste	Georgian	Tuscan (1.1)	Lezgin (1.6)
Sindhi[b]	10	Indo-European	Urban group	Georgian	Armenian (2.6)	Abhkasian (2.4)
Kashmiri Pandit	15	Indo-European	Upper caste	Georgian	Abhkasian (2)	Armenian (2.6)
Pathan[b]	15	Indo-European	Urban group	Georgian	Armenian (2.1)	Abhkasian (1.6)

We compute D(Onge, X; YRI, Y) where X is an Indian group shown above and Y is a West Eurasian group chosen from a panel of 42 groups including Europeans, Central Asians, Middle Easterners and Caucasian populations. We display the group with the highest D-statistic mean, 2nd highest D-statistic mean (Z-score for the difference between highest and 2nd highest group), and 3nd highest D-statistic mean (Z-score for the difference between the highest and 3nd highest). We consider $|Z| > 3$ to be statistically significant. [a] indicates samples from Metspalu et al (2011) and [b] indicates samples from HGDP.

Table D.3 (A): *D*-statistics differences: Madiga

Population (Y)	*n*	Country sampled from	Geographic Group	Mean difference	Z-score of difference
Lezgin	18	Caucasus	Caucasus	0.0001	1.1
Abhkasian	20	Caucasus	Caucasus	0.0001	1.2
Armenian	35	Armenia	Caucasus	0.0002	2.0
Kurd	6	Kazakhstan	Central Asia	0.0002	1.3
Chechen	20	Caucasus	Caucasus	0.0002	2.1
Cypriot	12	Cyprus	Europe	0.0003	2.2
Iranian	20	Iran	Central Asia	0.0003	2.8
Druze	42	Israel	Near East	0.0004	3.4
Syrian	16	Syria	Near East	0.0004	3.3
Adygei	17	Caucasus	Caucasus	0.0004	3.4
Tuscan	7	Italy	Europe	0.0004	3.0
North Ossetian	15	Russia	Caucasus	0.0004	3.5
TSI	87	Italy	Europe	0.0005	4.6
Lebanese	7	Lebanon	Near East	0.0005	3.1
Turk	19	Turkey	Near East	0.0005	4.7
Basque	24	France	Europe	0.0005	3.9
Kumyk	14	Russia	Caucasus	0.0005	4.5
CEU	110	United States	Europe	0.0005	5.0
Orcadian	15	United Kingdom	Europe	0.0005	3.9
Italian	12	Italy	Europe	0.0006	4.1
Jordanian	19	Jordania	Near East	0.0006	4.9
Hungarian	20	Hungary	Europe	0.0006	4.5
French	28	France	Europe	0.0006	5.0
Lithuanian	10	Lithuania	Europe	0.0006	3.9
Spaniard	12	Spain	Europe	0.0006	4.6
Bulgarian	13	Bulgaria	Europe	0.0006	5.0
Balkar	19	Caucasus	Caucasus	0.0006	5.8
Ukranian	20	Ukraine	Europe	0.0006	5.3
Palestinian	46	Israel	Near East	0.0007	6.1
Romanian	16	Romania	Europe	0.0007	5.6
Sardinian	28	Italy	Europe	0.0007	5.4
Saudi	19	Saudi Arabia	Near East	0.0007	5.2
Bedouin	45	Israel	Near East	0.0008	6.2
Belorussian	9	Belorussia	Europe	0.0008	5.5
Mordovian	15	Russia	Europe	0.0010	7.7
Russian	27	Russia	Europe	0.0011	9.0
Tajik	15	Tajikstan	Central Asia	0.0011	8.5
Yemenese	10	Yemen	Near East	0.0013	8.3
Turkmen	15	Turkmenistan	Central Asia	0.0016	11.3
Nogai	16	Russia	Caucasus	0.0017	13.3
Chuvash	17	Russia	Europe	0.0020	13.1
Uzbek	15	Uzbekstan	Central Asia	0.0032	19.7

We compare *D*(Onge, Madiga; YRI, Georgian) = 0.0335 ($Z = 16.7$) with *D*(Onge, Madiga; YRI, *Y*) where *Y* is any West Eurasian group chosen from a panel of 42 groups including Europeans, Central Asians, Middle Easterners and Caucasian populations. For each *Y*, we display the mean and Z-score of the difference with *D*(Onge, Madiga; YRI, Georgian).

Table D.3 (Continued) (B): D-statistics differences: Kashmiri Pandit

Population (X)	*n*	Country sampled from	Geographic Group	Mean difference	Z-score of difference
Abhkasian	20	Caucasus	Caucasus	0.0002	2.0
Lezgin	18	Caucasus	Caucasus	0.0003	2.2
Armenian	35	Armenia	Caucasus	0.0003	2.6
Cypriot	12	Cyprus	Europe	0.0004	2.8
Tuscan	7	Italy	Europe	0.0004	2.4
Chechen	20	Caucasus	Caucasus	0.0004	3.6
TSI	87	Italy	Europe	0.0005	4.9
Kurd	6	Kazakhstan	Central Asia	0.0005	3.0
Orcadian	15	United Kingdom	Europe	0.0005	3.6
CEU	110	United States	Europe	0.0005	4.8
Basque	24	France	Europe	0.0005	4.0
Italian	12	Italy	Europe	0.0005	3.8
Lithuanian	10	Lithuania	Europe	0.0005	3.5
French	28	France	Europe	0.0006	4.8
Druze	42	Israel	Near East	0.0006	5.4
Hungarian	20	Hungary	Europe	0.0006	5.0
Spaniard	12	Spain	Europe	0.0007	5.2
Bulgarian	13	Bulgaria	Europe	0.0008	5.8
Sardinian	28	Italy	Europe	0.0008	5.8
Adygei	17	Caucasus	Caucasus	0.0008	6.4
Ukranian	20	Ukraine	Europe	0.0008	6.3
Belorussian	9	Belorussia	Europe	0.0009	5.4
North Ossetian	15	Russia	Caucasus	0.0009	7.0
Syrian	16	Syria	Near East	0.0009	7.5
Romanian	16	Romania	Europe	0.0010	7.4
Turk	19	Turkey	Near East	0.0010	8.6
Iranian	20	Iran	Central Asia	0.0010	8.4
Balkar	19	Caucasus	Caucasus	0.0010	9.2
Lebanese	7	Lebanon	Near East	0.0010	6.4
Kumyk	14	Russia	Caucasus	0.0011	8.8
Jordanian	19	Jordania	Near East	0.0012	10.1
Palestinian	46	Israel	Near East	0.0013	11.1
Bedouin	45	Israel	Near East	0.0014	11.2
Mordovian	15	Russia	Europe	0.0014	10.6
Saudi	19	Saudi Arabia	Near East	0.0014	10.6
Russian	27	Russia	Europe	0.0015	12.2
Tajik	15	Tajikstan	Central Asia	0.0026	18.5
Yemenese	10	Yemen	Near East	0.0030	17.3
Nogai	16	Russia	Caucasus	0.0031	22.1
Turkmen	15	Turkmenistan	Central Asia	0.0031	20.2
Chuvash	17	Russia	Europe	0.0032	20.7
Uzbek	15	Uzbekstan	Central Asia	0.0057	31.6

We compare *D*(Onge, Kashmiri Pandit; YRI, Georgian) = 0.0627 (Z = 29.7) with *D*(Onge, Kashmiri Pandit; YRI, *Y*) where *Y* is any West Eurasian group chosen from a panel of 42 groups including Europeans, Central Asians, Middle Easterners and Caucasian groups. For each *Y*, we display mean and Z-score of the difference with *D*(Onge, Kashmiri Pandit; YRI, Georgian).

Table D.4: Ancestry estimates from *F_4 Ratio Estimation*

Population (X)	*n*	Language Group	Social/ caste group	§ANI Ancestry (Pop2 = Basque)	§ANI Ancestry (Pop2=Abhkasian)	§ANI ancestry (Reich 09)
Paniya	5	Dravidian	Tribal	16.7 ± 2.4	16.8 ± 2.1	22.5 ± 1.6
Palliyar	5	Dravidian	Tribal	21.2 ± 2.3	22.8 ± 2	29.1 ± 1.4
Kattunayakan	5	Dravidian	Tribal	24.6 ± 2.1	25.1 ± 1.9	30.8 ± 1.5
Palliyar[a]	5	Dravidian	Lower caste	24.2 ± 2.4	25.6 ± 2.1	31.4 ± 1.5
Madiga	13	Dravidian	Lower caste	32 ± 1.7	33.1 ± 1.5	40.6 ± 1.1
Mala	13	Dravidian	Lower caste	34.3 ± 1.7	35.8 ± 1.5	39.9 ± 1.1
Adi-Dravidar	5	Dravidian	Lower caste	34.7 ± 2	35.7 ± 1.7	40.9 ± 1.3
Hakkipikki[a]	4	Dravidian	Tribal	36.2 ± 2	35.4 ± 1.8	40.8 ± 1.4
Vedda	4	Indo-European	Tribal	36 ± 2.5	38 ± 2.2	41.3 ± 1.6
Kamsali	4	Dravidian	Lower caste	36.5 ± 2.1	38 ± 1.8	43.1 ± 1.4
Chenchu[a]	4	Dravidian	Tribal	37.2 ± 2.1	38.1 ± 1.9	43.4 ± 1.4
Chamar[a]	10	Indo-European	Tribal	38.7 ± 1.7	38.5 ± 1.5	43.1 ± 1.1
Chenchu	6	Dravidian	Tribal	39 ± 2.2	38.4 ± 2	41.7 ± 1.4
Bhil	17	Indo-European	Tribal	38.9 ± 1.6	39.3 ± 1.4	45.8 ± 1
Kallar	5	Dravidian	Lower caste	37.3 ± 2.1	39.4 ± 1.8	44.5 ± 1.3
Kallar[a]	8	Dravidian	Tribal	37.7 ± 1.8	40.4 ± 1.5	47.1 ± 1.1
Vysya	14	Dravidian	Middle caste	37.9 ± 1.8	41.2 ± 1.5	47.2 ± 1.1
Malai Kuravar	5	Dravidian	Tribal	38.8 ± 2.1	41.2 ± 1.9	46.8 ± 1.3
Satnami	3	Indo-European	Lower caste	40.7 ± 2.1	40.8 ± 1.9	43.4 ± 1.4
Kuruchiyan	5	Dravidian	Tribal	41.9 ± 1.9	43.2 ± 1.7	48.6 ± 1.2
Dushadh[a]	7	Indo-European	Lower caste	41 ± 1.8	42.8 ± 1.6	48.2 ± 1.2
Sch caste/ tribe[a]	6	Dravidian	Lower caste	40.5 ± 1.9	43.5 ± 1.6	48.8 ± 1.2
Mali	5	Dravidian	Lower caste	44 ± 2	43.3 ± 1.8	53.1 ± 1.2
Minicoy	4	Indo-European	Lower caste	42.9 ± 2	43.1 ± 1.7	48.9 ± 1.3
Gounder	5	Dravidian	Middle caste	42.9 ± 1.9	45.8 ± 1.7	51.8 ± 1.2
Lodi	5	Indo-European	Lower caste	43.1 ± 1.9	43.4 ± 1.7	49.4 ± 1.2
Naidu	4	Dravidian	Upper caste	43.2 ± 2	44.3 ± 1.8	48.5 ± 1.3
Velama	4	Dravidian	Upper caste	42.7 ± 2	46.3 ± 1.7	53.9 ± 1.3
Velama[a]	9	Dravidian	Upper caste	43.4 ± 1.7	45.3 ± 1.5	51.1 ± 1.1
Narikkuravar	5	Dravidian	Tribal	45 ± 2.2	46.1 ± 1.9	50.5 ± 1.5
Tharu	5	Indo-European	Tribal	43.6 ± 1.9	43.3 ± 1.7	50.5 ± 1.2
Dharkar[a]	11	Indo-European	Nomadic group	47.8 ± 1.5	47.3 ± 1.3	54.6 ± 1
Kanjar[a]	8	Indo-European	Nomadic group	48.2 ± 1.7	47.1 ± 1.5	53.5 ± 1.1
Muslim[a]	5	Indo-European	Religious group	49.4 ± 1.8	49.4 ± 1.5	55.1 ± 1.2
Srivastava	2	Indo-European	Upper caste	52.3 ± 2.5	51.6 ± 2.2	56.4 ± 1.5
Jain	5	Indo-European	Religious group	51.6 ± 1.9	52.1 ± 1.7	58 ± 1.2
Meghawal	5	Indo-European	Lower caste	53.6 ± 1.8	53.2 ± 1.6	58.2 ± 1.1
Kshatriya[a]	7	Indo-European	Upper caste	54.6 ± 1.6	53 ± 1.4	60.7 ± 0.9
Vaish	4	Indo-European	Upper caste	56.5 ± 1.7	54.5 ± 1.5	60.1 ± 1.2
Brahmin[a]	8	Indo-European	Upper caste	61.2 ± 1.4	57.8 ± 1.3	63.9 ± 0.9
Kshatriya	15	Indo-European	Upper caste	60.9 ± 1.3	58.4 ± 1.2	63.6 ± 0.8
Brahmin	10	Indo-European	Upper caste	62.8 ± 1.4	59.2 ± 1.3	64.5 ± 0.9
Sindhi[b]	10	Indo-European	Urban group	64.3 ± 1.3	62.7 ± 1.2	71.8 ± 0.8
Kashmiri Pandit	15	Indo-European	Upper caste	65.2 ± 1.3	63.8 ± 1.1	68.6 ± 0.8
Pathan[b]	15	Indo-European	Urban group	70.4 ± 1.2	67.9 ± 1	74.8 ± 0.7

We performed *F_4 Ratio Estimation* to estimate the proportion of ANI ancestry in Indians. Specifically, we use the following statistics: **§ANI ancestry (Pop2 = Basque)** = f_4(YRI, Basque; X, Onge)/ f_4(YRI, Basque; Georgians, Onge); **§ANI ancestry (Pop2 = Abhkasian)** = f_4(YRI, Abhkasian; X, Onge)/ f_4(YRI, Abhkasian; Georgians, Onge); **§ANI ancestry (Reich et al., 09)** = f_4(Adygei, Papuan; X, Onge)/ f_4(Adygei, Papuan; CEU, Onge). We computed standard errors using a Block Jackknife with a block size of 5cM. [a] indicates samples from Metspalu et al (2011) and [b] indicates samples from HGDP.

Table D.5. Dates of admixture using PCA loadings and one reference group

Pop	*n*	PCA based weights (*rolloff*)	PCA based weights (*ALDER*)	One reference (*ALDER*)
Indo-European rank 1	32	77 ± 6	70 ± 7	68 ± 12
Dravidian rank 1	33	121 ± 17	101 ± 17	105 ± 14

We performed *rolloff* and *ALDER* analysis using SNP loadings computed based on a PCA of Basque and all Indian cline groups (except the test groups). We also performed *ALDER* analysis using Basque as one reference group. To remove the effects of LD in the ancestral populations, we ignore bins corresponding to distance separation less than 0.5 cM: this threshold is set by *ALDER* after comparison of shared LD between Basque and the admixed groups.

Table D.6. Simulations to test bias in estimated dates of admixture for demographic parameters relevant to Indian groups

Group (X)	Sam-ples	ANI ancestry%	Simulated date based on point estimate from real data (gens)	Mean date estimate over 100 simulations
Brahmin	10	62.8 ± 1.4	65 ± 9	66
Mala	13	34.3 ± 1.7	96 ± 16	99
Pathan[b]	15	70.4 ± 1.2	73 ± 9	76
Dravidian rank 1	33	36.9 ± 1.5	121 ± 17	123
Indo-European rank 1	32	42.3 ± 1.4	77 ± 6	79

We simulated individuals of mixed European (CEU) and Asian (CHB) ancestry where we set sample size, ANI ancestry proportion, and the date of admixture to match the parameters in the real data for each Indian group (X). We performed *rolloff* using French and Han as the reference groups and computed the average admixture date (in generations) for 100 simulations.

Table D.7. Record of testing for consistency with simple ANI-ASI mixture.

Set size	Sets tested	Sets that are rank 1	Sets also passing *admixture graph*
3	7,770	3,692	1,152
4	25,425	5,152	860
5	19,239	1,293	90
6	1,852	30	1

Table D.8. Number of times each of 37 Indian groups is included in a passing set of groups

Population (X)	Linguistic affiliation	Traditional social status	Sampling Location	Set size = 3	Set size = 4	Set size = 5	Set size = 6	Set size = 7
Vysya	Dravidian	Middle caste	Andhra Pradesh	0	0	0	0	0
Sindhi[b]	Indo-European	Urban groups	Pakistan	2	0	0	0	0
Brahmin[a]	Indo-European	Upper caste	Uttar Pradesh	4	0	0	0	0
Chenchu	Dravidian	Tribal	Andhra Pradesh	8	0	0	0	0
Brahmin	Indo-European	Upper caste	Uttar Pradesh	15	0	0	0	0
Pathan[b]	Indo-European	Urban groups	Pakistan	19	2	0	0	0
Kshatriya	Indo-European	Upper caste	Uttar Pradesh	22	1	0	0	0
Mali	Dravidian	Lower caste	Lakshadweep	24	9	1	0	0
Kashmiri Pandit	Indo-European	Upper caste	Kashmir	33	10	0	0	0
Kshatriya[a]	Indo-European	Upper caste	Uttar Pradesh	40	13	0	0	0
Kanjar[a]	Indo-European	Nomadic group	Uttar Pradesh	48	23	0	0	0
Kallar[a]	Dravidian	Tribal	Tamil Nadu	50	81	20	1	0
Scheduled caste/ tribe[a]	Dravidian	Lower caste	Tamil Nadu	52	47	11	1	0
Gounder	Dravidian	Middle caste	Tamil Nadu	54	59	12	0	0
Velama[a]	Dravidian	Upper caste	Andhra Pradesh	62	71	11	0	0
Chamar[a]	Indo-European	Tribal	Uttar Pradesh	80	83	14	0	0
Dharkar[a]	Indo-European	Nomadic group	Uttar Pradesh	85	53	5	0	0
Kallar	Dravidian	Lower caste	Tamil Nadu	93	84	14	1	0
Narikkuravar	Dravidian	Nomadic group	Tamil Nadu	106	53	4	1	0
Paniya	Dravidian	Tribal	Kerela	107	81	3	0	0
Tharu	Indo-European	Tribal	Uttarkhand	118	91	10	0	0
Meghawal	Indo-European	Lower caste	Rajasthan	118	107	11	0	0
Bhil	Indo-European	Tribal	Gujarat	119	89	5	0	0
Madiga	Dravidian	Lower caste	Andhra Pradesh	123	112	8	0	0
Kattunayakan	Dravidian	Tribal	Kerela	123	82	10	0	0
Mala	Dravidian	Lower caste	Andhra Pradesh	123	167	24	0	0
Muslim[a]	Indo-European	Religious Group	Uttar Pradesh	124	87	5	0	0
Palliyar	Dravidian	Tribal	Tamil Nadu	124	99	15	0	0
Palliyar[a]	Dravidian	Lower caste	Tamil Nadu	128	81	8	0	0
Adi-Dravidar	Dravidian	Lower caste	Tamil Nadu	143	175	19	0	0
Dushadh[a]	Indo-European	Lower caste	Uttar Pradesh	158	169	15	0	0
Lodi	Indo-European	Lower caste	Uttar Pradesh	162	212	18	0	0
Malai Kuravar	Dravidian	Tribal	Tamil Nadu	164	195	29	1	0
Jain	Indo-European	Religious Group	Gujarat	169	193	32	0	0
Kuruchiyan	Dravidian	Tribal	Kerela	213	322	49	1	0
Onge	Jarawa-Onge	Hunter-gatherer	Andaman & Nicobar	215	311	53	0	0
Vedda	Indo-European	Tribal	Sri Lanka	227	278	44	0	0

Table D.9. Comparison of expected and observed weighted LD amplitudes for simulated data.

		Single wave				***Two waves***			
Europe %	**Ref. in *ALDER***	**Expected Amplitude x 10,000**	**Observed Amplitude x 10,000**	**Z**	α_{old}	**Expected Amplitude x 10,000**	**Observed Amplitude x 10,000**	**Z**	α_{old}
Simulation Set 1:									
30%	Basque	3.18 ± 0.13	3.30 ± 0.27	0.4	-0.8 ± 1.8	3.16 ± 0.15	1.76 ± 0.37	-3.7	11.3 ± 3.6
50%	Basque	1.83 ± 0.07	1.79 ± 0.23	-0.2	0.5 ± 3.1	1.81 ± 0.09	1.28 ± 0.12	-4.2	8.7 ± 2.3
70%	Basque	0.54 ± 0.03	0.56 ± 0.13	0.1	-0.6 ± 4.3	0.52 ± 0.04	0.11 ± 0.03	-14.2	36.6 ± 4.2
30%	Dai	0.60 ± 0.04	0.64 ± 0.14	0.3	-1.5 ± 4.5	0.58 ± 0.05	0.43 ± 0.12	-1.1	6.1 ± 5.6
50%	Dai	1.88 ± 0.09	1.70 ± 0.24	-0.8	2.6 ± 3.5	1.88 ± 0.10	1.16 ± 0.11	-4.1	11.9 ± 3.0
70%	Dai	3.04 ± 0.11	3.31 ± 0.31	0.8	-1.7 ± 2.0	3.07 ± 0.14	1.78 ± 0.23	-6.8	12.4 ± 2.7
Simulation Set 2:									
20%	Basque	3.36 ± 0.20	3.28 ± 0.23	-0.3	0.4 ± 1.5	3.45 ± 0.24	1.81 ± 0.23	-6.4	9.4 ± 1.5
30%	Basque	3.13 ± 0.18	3.37 ± 0.31	0.7	-1.7 ± 2.6	3.32 ± 0.25	1.88 ± 0.28	-3.6	10.9 ± 3.1
40%	Basque	2.55 ± 0.16	2.85 ± 0.31	0.9	-2.8 ± 3.2	2.68 ± 0.25	1.50 ± 0.24	-4.0	13.2 ± 3.4
20%	Dai	0.22 ± 0.03	0.10 ± 0.07	-1.7	10.8 ± 6.8	0.18 ± 0.04	0.22 ± 0.44	0.1	-3.3 ± 31.7
30%	Dai	0.63 ± 0.05	0.59 ± 0.09	-0.4	1.4 ± 3.5	0.51 ± 0.08	0.29 ± 0.09	-2.1	11.1 ± 5.2
40%	Dai	1.21 ± 0.09	1.29 ± 0.19	0.4	-1.5 ± 3.4	1.11 ± 0.16	0.92 ± 0.12	-1.2	4.5 ± 3.7

Standard errors shown are based on jackknife estimates from a single simulation (not standard errors from averaging over multiple simulations). To infer the statistical uncertainty of (Observed - Expected) amplitude, we use a weighted block jackknife dropping each chromosome in turn and repeating the entire procedure. This produces a standard error and allows us to test whether the difference is consistent with zero ($|Z| < 3$).

References

1. 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 Genetics 7,e1001373.

2. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. Genetics 192, 1065-1093.

3. Busing, F., Meijer, E., and Leeden, R. (1999). Delete-m Jackknife for Unequal m. Statistics and Computing 9, 3-8.

4. Moorjani, P., Patterson, N., Loh, P.-R., Lipson, M., Kisfali, P., Melegh, B.I., Bonin, M., Kádaši, Ľ., Rieß, O., Berger, B., et al. (2013). Reconstructing Roma history from genome-wide data. PloS one 8, e58633.

5. Loh, PR., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. (2013). Infering Admixture Histories of Human Populations using Weighted Linkage Disequilibrium. Genetics 193, 1233-1254.

6. Reich, D., Thangaraj, K., Patterson, N., Price, A., and Singh, L. (2009). Reconstructing Indian population history. Nature 461, 489-494.

7. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., and Mesa, N. (2012). Reconstructing Native American population history. Nature 488, 370-374.

8. Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337.

9. Barik, S., Sahani, R., Prasad, B., Endicott, P., Metspalu, M., Sarkar, B., Bhattacharya, S., Annapoorna, P., Sreenath, J., and Sun, D. (2008). Detailed mtDNA genotypes permit a reassessment of the settlement and population structure of the Andaman Islands. American Journal of Physical Anthropology 136, 19-27.

10. Shah, A.M., Tamang, R., Moorjani, P., Rani, D.S., Govindaraj, P., Kulkarni, G., Bhattacharya, T., Mustak, M.S., Bhaskar, L., and Reddy, A.G. (2011). Indian siddis: African descendants with Indian admixture. The American Journal of Human Genetics 89, 154-161.

11. Metspalu, M., Romero, I.G., Yunusbayev, B., Chaubey, G., Mallick, C.B., Hudjashov, G., Nelis, M., Mägi, R., Metspalu, E., and Remm, M. (2011). Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. The American Journal of Human Genetics 89, 731-744.