Statistical methods for detecting admixture

Pongsakorn Wangkumhang & Garrett Hellenthal

University College London Genetics Institute (UGI), Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

The increasing availability of large-scale autosomal genetic variation data sampled from world-wide geographic areas, coupled with advances in the statistical methodology to analyse these data, is showcasing the power of DNA as a major tool to gain insights into the demographic history of humans and other organisms. Here we review statistical techniques that shed light on a specific aspect of demography: the detection and description of admixture events where two or more genetically distinct groups intermixed at one or more times in the past. In particular we give an overview of some of the widely-used methods to identify and describe admixture events using autosomal DNA from unrelated individuals, with a particular focus on analysing biallelic Single-Nucleotide-Polymorphsim (SNP) markers.

While Y-chromosome and mitochondrial (mtDNA) have been extremely valuable in studying sex-biased admixture, where one of the admixing groups contributes a disproportionate number of males or females [1], current widely-used approaches focus on analysing autosomal DNA as it contains many thousands of times more (sex-averaged) information than Y/mtDNA. As examples, we apply some of these approaches to two sets of simulated data from [2] containing 474,491 autosomal SNPs (Figure 1). Each set consists of 20 simulated individuals, descending from a pulse of admixture between two sources occurring 30 generations ago. These sources are African and European in the first set, contributing ≈80% and ≈20% of the DNA, respectively, while the sources are Central-South-Asian and European in the second set, each contributing ≈50% of the DNA. Following the procedure in [3], these simulations used DNA from 21 present-day Yoruban individuals from Nigeria, 21 present-day Brahui individuals from Pakistan and 28 present-day individuals from France as the African, Central-South-Asian and European admixing sources, respectively. In analyses below, we also include surrogate individuals representing each of these continental admixing sources, in particular using the genomes of 22 Mandenka from Africa, 21 Balochi from Central South Asia, and 23 British/Irish from Europe.

Signatures of admixture

Two of the most widely-used techniques for describing population structure and demography are principal components analysis (PCA) and clustering algorithms. While the genetic patterns captured by each approach may be attributable to factors other than admixture, they can each highlight individuals who descend from multiple intermixing source groups.

Data dimension reduction techniques

PCA is an algebraic technique that applies eigenvector decomposition to reduce high-dimensional data to a small number of independent variables, termed principal components, that explain a large proportion of variation in the complete data. Typically PCA is applied to the genome-wide genotype data of multiple individuals, and the top few principal components are plotted to visualize genetic distance among the individuals [4, 5, 6, 7, 8, 9]. For example, Figure 1b plots the first two principal components from the PCA program EIGENSTRAT [6] applied to our simulated data and the surrogate individuals. Note that surrogate individuals from the three continental groups fall into different corners of the plot, while each set of admixed individuals fall between the two continental groups comprising their ancestry in a manner roughly consistent with admixture proportions. However, caution is strongly warranted against concluding admixture wherever such a pattern is observed, as PCA projections can depend on sample size, SNP ascertainment, and features of demography besides admixture [9, 10].

To better interpret these patterns, recent approaches have related these concepts to isolation-by-distance models that assume genetic similarity decreases with geographic distance between samples. For example, SpaceMix [11] highlights populations whose allele frequencies are more correlated than expected when assuming genetic similarity decays exponentially with distance, while EEMS [12] relates genetic similarity to the distance among individuals' geographic locations (e.g. birthplace or sampling information) assuming migration occurs between neighboring demes via a stepping stone model [13]. Each approach then attempts to identify genetic associations beyond that explained by these simple distance models, which may indicate long-distance admixture.

Clustering algorithms

Clustering algorithms attempt to classify individuals into a discrete number of groups based on associations among their genetic patterns. Popular algorithms for doing so include STRUCTURE [14], ADMIXTURE [15], FRAPPE [16] and related models [17, 18]. While these techniques assume loci are unlinked, STRUCTURE version 2.0 [19] models linkage among loci caused by the inheritance of blocks of DNA from different admixing sources (as in Fig 1a), while still ignoring correlations among tightly linked loci (though see fineSTRUCTURE [20]). Typically users select a number of clusters K (though K can also be estimated, e.g. [17, 20]), and many of these programs then determine the proportion of each individual's DNA derived from K inferred clusters. Therefore, it can be tempting to interpret the K inferred clusters as K ancestral source groups that potentially intermixed, so that individuals whose DNA is assigned to multiple clusters descend from historical admixture events. Under this interpretation, these programs also estimate the proportion of DNA contributed by each source. Figure 1c provides results from ADMIXTURE applied to the simulated and surrogate data using $K = 2 - 5$. Here ADMIXTURE accurately describes sources and proportions of admixture in the two simulated groups at $K = 3$, but does not always characterize this admixture well at other values of K, with cross-validation choosing $K = 2$ as the best-fitting value. Therefore, as in PCA, multiple disparate demographic histories can lead to identical clustering [14, 21], warranting caution against concluding admixture when individuals are classified as mixtures of different clusters or – conversely – concluding no admixture when individuals are assigned to a single cluster.

Identifying admixture

f statistics for admixture testing

If population C descends from an admixture event between two populations A and B, the allele frequency p_C of a variant at a locus (e.g. a SNP) in C typically should fall between the frequencies p_A, p_B of that variant in populations A and B, respectively. Therefore, to test for admixture, Reich et al [22] propose the $f3$ statistic, defined as an unbiased estimator for $(p_c - p_A)(p_c - p_B)$ averaged across all loci. This statistic is significantly negative if population C descends from an admixture event between two sources related to A and B, i.e. if p_C falls between p_A and p_B, with significance assessed using jack-knife re-sampling based on dividing the genome into independent regions (e.g. chromosomes). Applying ADMIXTOOLS to calculate $f3$ statistics for the two simulated populations of Figure 1 using the given surrogates finds strong evidence of admixture in both (Z-scores < -30). However, the authors note that the $f3$ statistic may no longer be negative if C has experienced a high degree of drift following admixture, e.g. due to a bottleneck event, as this can cause p_C to no longer often fall between p_A and p_B [23]. A related test is the $f4$ statistic, which estimates the average of $(p_A - p_B)(p_C - p_D)$ across loci using the sample allele frequencies from four populations A-D [22] to formally test whether (A, B) and (C, D) form clades. If the $f4$ statistic is significantly positive (respectively negative), then A (respectively B) is more related ancestrally to C and/or B (respectively A) is more related to D than expected, possibly indicating admixture or incorrect clades. Ratios of $f4$ statistics can also be used to infer proportions of admixture from each source, assuming the tree topology among populations is known [22, 23].

Tree-based inference

Expanding upon these ideas, a topology detailing the order of splits among a set of n populations can be constructed based on correlations among those populations' allele frequencies. For example, qpGraph [22, 23] scores how well observed f statistics fit those predicted by a user-specified bifurcating tree relating the n populations, which may contain multiple migration events. The authors caution that n should be small in practice. MixMapper [24] extends qpGraph to a larger number of populations, by first inferring which populations are unadmixed using $f3$ statistics, then building a bifurcating tree using these putatively unadmixed populations, and finally adding admixed populations onto this tree. TREEMIX [25] is a related model that builds bifurcating trees relating a large number of groups and then adds links between branches to indicate admixture. TREEMIX uses a multivariate normal distribution to relate observed allele frequencies among populations, incorporating work by [26, 27], with the mean and covariance of this distribution defined by the (unknown) tree branch lengths (measuring drift) and admixture links among populations. Figure 1d shows the inferred tree topology and admixture events when applying TREEMIX to the simulated data and surrogate individuals, correctly characterising the admixture in both cases. While useful for describing relations among populations and unearthing migration events, a major limitation of these approaches is that the unknown complexity of genetic relationships among large numbers of populations necessitates exploring an enormous search space encompassing the possible tree topologies and potential migration events. Therefore, while these approaches can provide solutions that fit patterns in allele frequency correlation across populations, it is difficult to assess the extent to which other equally-likely solutions exist [28].

Dating admixture

The block-like manner in which autosomal DNA is inherited potentially enables identifying when admixture occurred. For convenience, many approaches assume two or more ancestral sources intermixed over a short time interval (e.g. one generation), often referred to as a "pulse" of admixture, after which individuals in the admixed population randomly mate for subsequent generations. In this setting, within the admixed genomes r generations after this admixture event, the probability that two loci were inherited from the same ancestor at the time of admixture decays exponentially with rate r per Morgan [19]. Thus fitting an exponential rate of decay to the association between loci attributable to admixture can be used to determine r. Violations of this simplistic model, e.g. continuous migration, can be diagnosed by studying whether a single exponential decay function is not a good fit to the data or by mathematically formulating the expected pattern of decay under more complex models. In contrast to e.g. $f3$ statistics, typically only admixture within a few hundred generations can be detected using these approaches, since beyond this segment sizes will have decreased to levels not distinguishable from levels of noise. Another potential issue is how well an exponential function captures certain admixture scenarios [29].

Techniques to infer local ancestry tracts

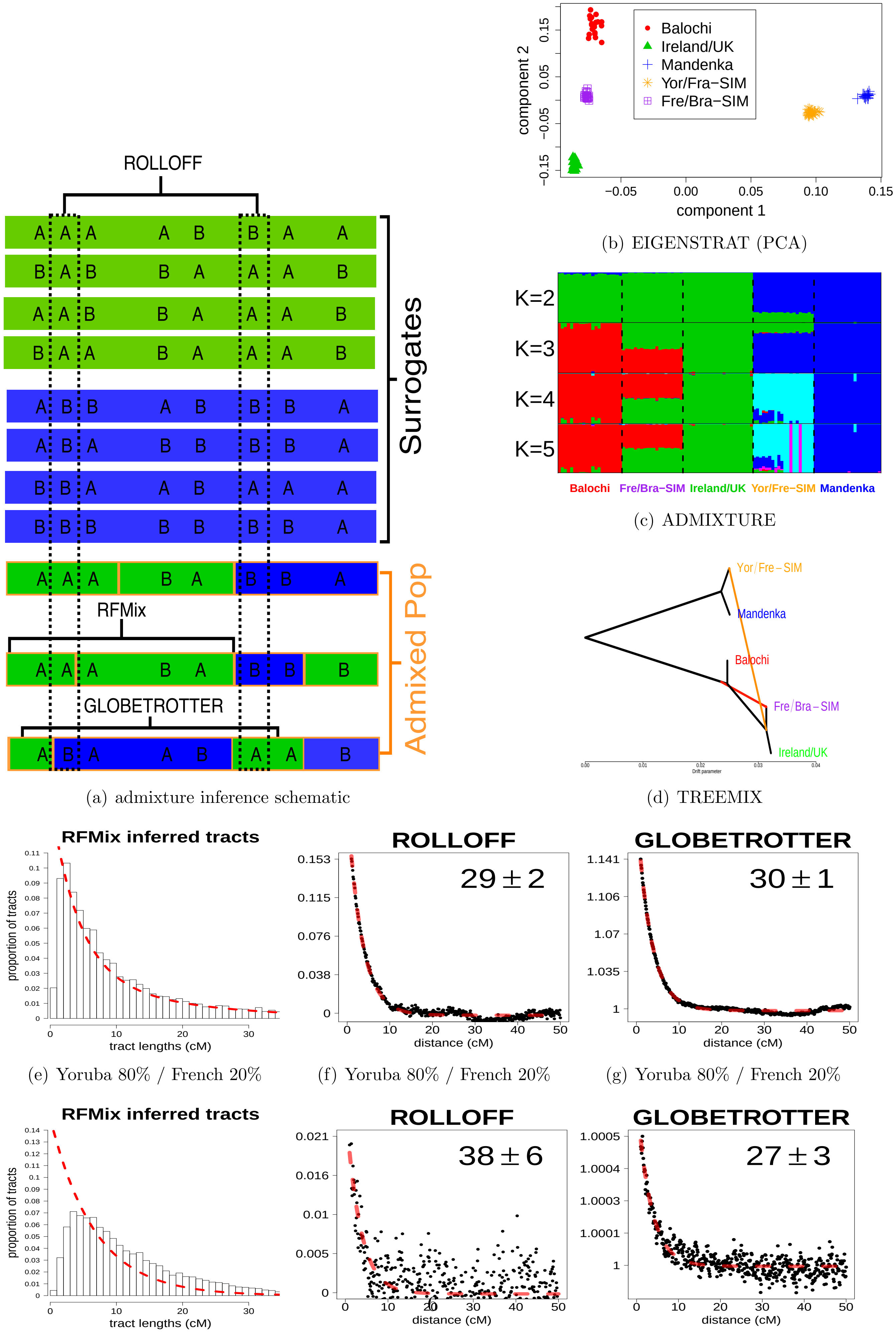
In theory, within each admixed individual contiguous segments (or "tracts") of DNA inherited from the original admixing sources can be identified by matching the admixed individual's genetic variation patterns to that of a set of surrogates meant to represent the source groups (Figure 1a). The sizes of these inferred tracts are then compared to those expected under different models of migration, to determine the best-fitting date(s) of admixture and the proportions of ancestry inherited from each admixing source [30, 31]. Several different algorithms have been proposed to do this matching (e.g. [32, 33, 3, 34, 35, 36, 37, 38, 39, 40]) that have different advantages, such as the number of putative admixing sources allowed, whether the data must be pre-phased, whether tightly linked haplotype information is directly modeled, etc, with the details of many summarized in [38].

As admixing sources must be distinguishable using small numbers of SNPs in local region, these approaches are only applicable to recent admixture between genetically diverged sources, such as Latin and African American human data, where tracts are both longer and easier to classify. As an illustration of this, Figure 1e,h summarises results from applying RFMix [39], which uses linear discriminant analysis to match segments in admixed individuals to those from reference populations, to the simulated and surrogate data from [2]. There is much better agreement between the expected exponential decay and the observed tract length distribution inferred by RFMix for the Yoruba-French simulation relative to the French-Brahui simulation, because the two admixing sources are considerably more genetically similar in the latter.

Techniques using linkage disequilibrium decay patterns

Other approaches avoid directly inferring local tracts of DNA by instead measuring how the association among loci, i.e. linkage disequilibrum (LD), attributable to admixture decays with increasing genetic distance. For example, ROLLOFF [41, 42, 23] and ALDER [43] measure the decay in LD versus genetic distance among pairs of biallelic loci (e.g. SNPs), weighted by each locus's ability to distinguish between surrogates of the original admixing populations. The weighted covariance (or correlation) between pairs of SNPs in the admixed populations should decay at rate r per Morgan, and the amplitude of these decay curves can be used to infer the proportion of admixture [43]. To potentially increase power, GLOBETROTTER [2] instead measures the association among pairs of haplotype segments in the admixed individuals, rather than pairs of SNPs. Here each haplotype segment is defined by the surrogate population it is most closely related to ancestrally, as inferred using CHROMOPAINTER [20] on pre-phased data. ROLLOFF and ALDER infer the best surrogates for each source by finding the best model fit out of all possible pairings of available surrogates, while GLOBETROTTER infers the genetic make-up of each source as a mixture of DNA from all potential surrogate groups (i.e. without requiring pre-specification of surrogates). Both can identify multiple admixture events at different times [44, 2], and GLOBETROTTER can also infer whether >2 sources intermixed at approximately the same time.

Typically with these techniques, inferring proportions of admixture from each source is more challenging than inferring dates of admixture, with the former often suffering from a lack of identifiability (e.g. surrogate populations may themselves be admixed to an unknown degree). Furthermore, in the cases of continuous admixture or migration at several different times, inferred dates from these approaches may be biased towards the most recent date [43, 2]. Nonetheless, decay of LD due to admixture can still be usefully modeled by these approaches in far more subtle cases relative to approaches that directly identify tracts. For example, note the considerably more accurate modeling of the true date in the French-Brahui simulation (Figure 1h-j), with GLOBETROTTER showing more precision than ROLLOFF in these simulations.



(a) admixture inference schematic

(b) EIGENSTRAT (PCA)

(c) ADMIXTURE

(d) TREEMIX

(e) Yoruba 80% / French 20%

(f) Yoruba 80% / French 20%

(g) Yoruba 80% / French 20%

(h) Brahui 50% / French 50%

(i) Brahui 50% / French 50%

(j) Brahui 50% / French 50%

Figure 1. Identifying admixture in simulated data from [2]. (a) Populations represented by blue/green surrogate groups mix r generations in the past, generating admixed haplotypes at bottom. A and B denote allele types at each biallelic SNP; orange bars separate segments in admixed chromosomes that exactly match a region of a depicted surrogate haplotype. (b) EIGENSTRAT (each dot an individual), (c) ADMIXTURE for cluster numbers $K = 2 - 5$ (columns = individuals, colors = clusters) and (d) TREEMIX (black lines = inferred topology; red/orange lines = migration edges) applied to allele frequency data from individuals simulated as mixtures of 50% French + 50% Brahui ("Fre/Bra-SIM") or 80% Yoruba + 20% French ("Yor/Fre-SIM"), also incorporating (Balochi,Ireland/UK,Mandenka) as surrogates for (Brahui,French,Yoruba), respectively. (e)-(j) RFMix, ROLLOFF, and GLOBETROTTER applied to date admixture in these simulations, with each approach intuitively explained in (a) as measuring (e,h) the distribution of tract lengths matched to the same surrogate group, or the decay of LD (black dots) versus genetic distance between (f,i) SNPs differentiated between the two surrogates or (g,j) haplotype blocks matched to a surrogate (Ireland). Dashed red lines give the expected decay for the true proportions and true admixture date of 30 generations ago, and numbers in the top right of (f,g,i,j) give inferred dates and standard errors.

Conclusion and perspectives

Statistical advances have rapidly increased the amount of admixture information we can extract out of large-scale autosomal variation data, with admixture events within the last 4,000 years now characterized in most global populations [43, 2]. However, one notable limitation is that most approaches rely on using surrogates for the original (unknown) admixing sources, and it is unclear how accurate these surrogates may be. For example, often modern-day samples are used as surrogates despite themselves having recent admixture from other sources. While the rapid increase of reliable DNA extractions from ancient human remains, i.e. aDNA (e.g. [45]), eventually may provide more accurate surrogates to the original (possibly unadmixed) sources, it is unclear from their current sparcity whether enough representative aDNA can be obtained for many admixture events of interest. Such aDNA also opens the possibility for unearthing details of admixture further back in time, e.g. dating admixture within aDNA samples [46]. In addition, the increasing availability of large-scale genetic variation resources from modern individuals will enable more precise inference, such as better classification of continuous versus pulses of admixture and better ability to detect subtle admixture between genetically similar sources [47]. Furthermore forthcoming large-scale sequencing datasets that capture rare genetic variants will enable better identification of recent shared ancestry among individuals.

Finally, since the study of demographic processes such as admixture affects the whole genomes of individuals, one promising avenue of future work is using these detailed inferences as a null model when testing specific regions of the genome for evidence of selection. Recent examples include such extensions for PCA approaches [48] and for MixMapper [49], as well as using local admixture tract inference approaches to identify whether admixed populations contain more ancestry from particular sources than average in specific genomic regions [50, 40]. In this way these models not only enable understanding of our history, but may also point towards genetic regions facilitating our adaptation to disease pressures in a complimentary manner to on-going genome-wide association studies.

Acknowledgements

GH is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 098386/Z/12/Z) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. PW is supported by a Royal Thai Government Scholarship.

References

[1] M.A. Jobling, M.E. Hurles, and C. Tyler-Smith. *Human Evolutionary Genetics: Origins, Peoples and Disease.* Garland Science, Abingdon and New York, 2004.

[2] G. Hellenthal, G.B.J. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, and S. Myers. A genetic atlas of human admixture history. *Science*, 343:747–751, 2014.

[3] A.L. Price, A. Tandon, N. Patterson, K.C. Barnes, N. Rafaels, I. Ruczinski, T.H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genetics*, 5(6):e1000519, 2009.

[4] Menozzi, P. and Piazza, A. and Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science*, 201:786–792, 1978.

[5] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Gen*, 2(12):e190, 2006.

[6] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:504–509, 2006.

[7] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergman, M.R. Nelson, M. Stephens, and C.D. Bustamante. Genes mirror geography within Europe. *Nature*, 456:98–101, 2008.

[8] Lao, O. and Lu, T.T. and Nothnagel, M. and Junge, O. and Freitag-Wolf, S. and Caliebe, A. and Balascakova, M. and Bertranpetit, J. and Bindoff, L.A. and Comas, D. and Holmlund, G. and Kouvatsi, A. and Macek, M. and Mollet, I. and Parson, W. and Palo, J. and Ploski, R. and Sajantila, A. and Tagliabracci, A. and Gether, U. and Werge, T. and Rivadeneira, F. and Hofman, A. and Uitterlinden, A.G. and Gieger, C. and Wichmann, H.E. and Ruther, A. and Schreiber, S. and Becker, C. and Nurnberg, P. and Nelson, M.R. and Krawczak, M. and Kayser, M. Correlation between genetic and geographic structure in Europe. *Current Biology*, 18(16):1241–1248, 2008.

[9] G. McVean. A genealogical interpretation of principal components. *PLoS Genet*, 5(10):e1000686, 2009.

[10] Novembre, J. and Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40:646–649, 2008.

[11] Bradburd, G.S. and Ralph, P.L. and Coop, G.M. A spatial framework for understanding population structure and admixture. *PLoS Genet*, 12(1):e1005703, 2016.

[12] Petkova, D. and Novembre, J. and Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genetics*, 48(1):94–100, 2016.

[13] Kimura, M. and Weiss, G.H. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49:561–576, 1964.

[14] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotypes data. *Genetics*, 155:945–959, 2000.

[15] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

[16] H. Tang, J. Peng, P. Wang, and N. Risch. Estimation of Individual Admixture: Analytical and Study Design Considerations. *Genetic Epidemiology*, 28:289–301, 2005.

[17] Huelsenbeck, J.P. and Andolfatto, P. and Huelsenbeck, E.T. Structurama: Bayesian inference of population structure. *Evol Bioinform Online*, 7:55–59, 2011.

[18] Raj, A. and Stephens, M. and Pritchard, J.K. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, 197(2):573–589, 2014.

[19] D. Falush, M. Stephens, and J.K. Pritchard. Inference of population structure from multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

[20] D.J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453, 2012.

[21] L. van Dorp, D. Balding, S. Myers, L. Pagani, C. Tyler-Smith, E. Bekele, A. Tarekegn, M.G. Thomas, N. Bradman, and G. Hellenthal. Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLoS Genetics*, 11(8):e1005397, 2015.

[22] D. Reich, K. Thangaraj, N. Patterson, A.L. Price, and L. Singh. Reconstructing Indian population history. *Nature*, 461:489–494, 2009.

[23] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient Admixture in Human History. *Genetics*, 192(3):1065–1093, 2012.

[24] Lipson, M. and Loh, P.R. and Levin, A. and Reich, D. and Patterson, N. and Berger, B. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Mol Biol Evol*, 30(8):1788–1802, 2013.

[25] J.K. Pickrell and J.K. Pritchard. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet*, 8:e1002967, 2012.

[26] Cavalli-Sforza, L.L. and Edwards, A.W. Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet*, 19:233–257, 1967.

[27] Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet*, 25:471–492, 1973.

[28] Mathieson, I. and Alpaslan-Roodenberg, S. and Posth, C. and Szcsnyi-Nagy, A. and Rohland, N. and Mallick, S. and Olalde, I. and Broomandkhoshbacht, N. and Candilio, F. and Cheronet, O. and Fernandes, D. and Ferry, M. and Gamarra, B. and Fortes, G.G. and Haak, W. and Harney, E. and Jones, E. and Keating, D. and Krause-Kyora, B. and Kucukkalipci, I. and Michel, M. and Mittnik, A. and Ngele, K. and Novak, M. and Oppenheimer, J. and Patterson, N. and Pfrengle, S. and Sirak, K. and Stewardson, K. and Vai, S. and Alexandrov, S. and Alt, K.W. and Andreescu, R. and Antonovi, D. and Ash, A. and Atanassova, N. and Bacvarov, K. and Gusztv, M.B. and Bocherens, H. and Bolus, M. and Boronean, A. and Boyadzhiev, Y. and Budnik, A. and Burmaz, J. and Chohadzhiev, S. and Conard, N.J. and Cottiaux, R. and uka, M. and Cupillard, C. and Drucker, D.G. and Elenski, N. and Francken, M. and Galabova, B. and Ganetsovski, G. and Gly, B. and Hajdu, T. and Handzhyiska, V. and Harvati, K. and Higham, T. and Iliev, S. and Jankovi, I. and Karavani, I. and Kennett, D.J. and Komo, D. and Kozak, A. and Labuda, D. and

Lari, M. and Lazar, C. and Leppek, M. and Leshtakov, K. and Vetro, D.L. and Los, D. and Lozanov, I. and Malina, M. and Martini, F. and McSweeney, K. and Meller, H. and Menui, M. and Mirea, P. and Moiseyev, V. and Petrova, V. and Price, T.D. and Simalcsik, A. and Sineo, L. and laus, M. and Slavchev, V. and Stanev, P. and Starovi, A. and Szeniczey, T. and Talamo, S. and Teschler-Nicola, M. and Thevenet, C. and Valchev, I. and Valentin, F. and Vasilyev, S. and Veljanovska, F. and Venelinova, S. and Veselovskaya, E. and Viola, B. and Virag, C. and Zaninovi, J. and Zuner, S. and Stockhammer, P.W. and Catalano, G. and Krau, R. and Caramelli, D. and Zaria, G. and Gaydarska, B. and Lillie, M. and Nikitin, A.G. and Potekhina, I. and Papathanasiou, A. and Bori, D. and Bonsall, C. and Krause, J. and Pinhasi, R. and Reich, D. The genomic history of southeastern Europe. *Nature*, 555:197–203, 2018.

[29] M. Liang and R. Nielsen. The lengths of admixture tracts. *Genetics*, 197:953–967, 2014.

[30] J. E. Pool and R. Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181:711–719, 2009.

[31] Gravel, G. Population genetics models of local ancestry. *Genetics*, 191:607–619, 2012.

[32] H. Tang, M. Coram, P. Wang, X. Xhu, and N. Risch. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet*, 79:1–12, 2006.

[33] S. Sriram Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating Local Ancestry in Admixed Populations. *American Journal of Human Genetics*, 82(2):290–303, 2008.

[34] Bryc, K. and Auton, A. and Nelson, M.R. and Oksenberg, J. R. and Hauser, S.L. and Williams, S. and Froment, A. and Jean-Marie Bodo, J.M. and Charles Wambebe, C. and Tishkoff, S.A. and Bustamante, C.D. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA*, 107(2):786–791, 2010.

[35] Baran, Y. and Pasaniuc, B. and Sankararaman, S. and Torgerson, D.G. and Gignoux, C. and Eng, C. and Rodriguez-Cintron, W. and Chapela, R. and Ford, J.G. and Avila, P.C. and Rodriguez-Santana, J. and Burchard, E.G. and Halperin, E. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10):1359–67, 2012.

[36] Brisbin, A. and Bryc, K. and Byrnes, J. and Zakharia, F. and Omberg, L. and Degenhardt, J. and Reyonlds, A. and Ostrer, H. and Mezey, J.G. and Bustamante, C.D. PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum Biol*, 84(4):343–364, 2012.

[37] Omberg, L. and Salit, J. and Hackett, N. and Fuller, J. and Matthew, R. and Chouchane, L. and Rodriguez-Flores, J.L. and Bustamante, C. and Crystal, R.G. and Mezey, J.G. Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC Genetics*, 13:49, 2012.

[38] Churchhouse, C. and Marchini, J. Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic Epidemiology*, 37(1):1–12, 2013.

[39] Maples, B.K. and Gravel, S. and Kenny, E.E. and Bustamante, C.D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet*, 93(2):278–288, 2013.

[40] Y. Guan. Detecting structure of haplotypes and local ancestry. *Genetics*, 196:625–642, 2014.

[41] P. Moorjani, N. Patterson, J.N. Hirschhorn, A. Keinan, L. Hao, G. Atzmon, E. Burns, H. Ostrer, A.L. Price, and D. Reich. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics*, 7(4):e1001373, 2011.

[42] P. Moorjani, N. Patterson, P.R. Loh, M. Lipson, P. Kisfali, B.I. Melegh, M. Bonin, L. Kadasi, O. Riess, B. Berger, D. Reich, and B. Melegh. Reconstructing Roma history from genome-wide data. *PLoS ONE*, 8(3):e58633, 2013.

[43] P.R. Loh, M. Lipson, N. Patterson, P. Moorjani, J.K. Pickrell, D. Reich, and B. Berger. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, 193(4):1233–1254, 2013.

[44] J.K. Pickrell, N. Patterson, P.R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, and D. Reich. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci USA*, 111(7):2632–7, 2014.

[45] I. Lazaridis, D. Nadel, G. Rollefson, D.C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424, 2016.

[46] Lipson, M. and Szecsenyi-Nagy, A. and Mallick, S. and Posa, A. and Stgmr, B. and Keerl, V. and Rohland, N. and Stewardson, K. and Ferry, M. and Michel, M. and Oppenheimer, J. and Broomandkhoshbacht, N. and Harney, E. and Nordenfelt, S. and Llamas, B. and Gusztv Mende, B. and Khler, K. and Oross, K. and Bondr, M. and Marton, T. and Oszts, A. and Jakucs, J. and Paluch, T. and Horvth, F. and Csengeri, P. and Kos, J. and Sebk, K. and Anders, A. and Raczky, P. and Regenye, J. and Barna, J.P. and Fbin, S. and Serlegi, G. and Toldi, Z. and Gyngyvr Nagy, E. and Dani, J. and Molnr, E. and Plfi, G. and Mrk, L. and Melegh, B. and Bnfai, Z. and Domborczki, L. and Fernndez-Eraso, J. and Antonio Mujika-Alustiza, J. and Alonso Fernndez, C. and Jimnez Echevarra, J. and Bollongino, R. and Orschiedt, J. and Schierhold, K. and Meller, H. and Cooper, A. and Burger, J. and Bnffy, E. and Alt, K.W. and Lalueza-Fox, C. and Haak, W. and Reich, D. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*, 551:368–372, 2017.

[47] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E.C. Royrvik, B. Cunliffe, Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, D.J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, and W. Bodmer. The fine scale genetic structure of the British population. *Nature*, 519:309–314, 2015.

[48] K.J. Galinsky, G. Bhatia, P.R. Loh, S. Georgiev, S. Mukherjee, N.J. Patterson, and A.L. Price. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet*, 98(3):456–72, 2016.

[49] Racimo, F. and Berg, J.J. and Pickrell, J.K. Detecting polygenic adaptation in admixture graphs. *Genetics*, 208:1565–84, 2018.

[50] H. Tang, S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Clintron, E. Gonzalez, and N.J. Risch. Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. *Amer J Hum Genet*, 81(3):626–633, 2007.