A Genetic Atlas of Human Admixture History

Garrett Hellenthal,1 George B. J. Busby,2 Gavin Band,3 James F. Wilson,4 Cristian Capelli,2 Daniel Falush,* Simon Myers3,6†

Modern genetic data combined with appropriate statistical methods have the potential to contribute substantially to our understanding of human history. We have developed an approach that exploits the genomic structure of admixed populations to date and characterize historical mixture events at fine scales. We used this to produce an atlas of worldwide human admixture history, constructed by using genetic data alone and encompassing over 100 events occurring over the past 4000 years. We identified events whose dates and participants suggest they describe genetic impacts of the Mongol empire, Arab slave trade, Bantu expansion, first millennium CE migrations in Eastern Europe, and European colonialism, as well as unrecorded events, revealing admixture to be an almost universal force shaping human populations.

**Fig. 1. Ancestry painting and admixture analysis of simulated admixture.** (A) A simulated event 30 generations ago between Brahui (80%, red) and Yoruba (20%, yellow) resulted in admixed individuals having DNA segments from each source (bottom). The true sources are then treated as unsampled, cM, centimorgan. (B) CHROMOPAINTER’s painting of the same region (yellow, Africa; green, America; red, Central-South Asia; blue, East Asia; cyan, Europe; pink, Near East; black, Oceania), showing haplotypic segments (“chunks”) shared with these groups. Our model fitting narrows the donor set largely to Central-South Asia and Africa, generating a “cleaned” painting. (C) Coancestry curves (black line) show relative probability of jointly copying two chunks from red (Bolochi; closest surrogate pop ≤ 0.003 with Brahui) and/or yellow (Mandenka; FST = 0.009 with Yoruba) donors, at varying genetic distances. The curves closely fit an exponential decay (green line) with a rate of 30 generations (95% CI: 27 to 33). The positive slope for the Bolochi-Mandenka curve (middle) implies that these donors represent different admixing sources. (D) GLOBETROTTER’s source inference, with black diamonds indicating sampled populations with greatest similarity (FST ≤ 0.001 over minimum) to true sources, white circles other sampled populations. Red and yellow circles, with areas summing to 20% and 80%, respectively, show inferred haplotypic makeup of the two admixing sources.
polymorphisms (SNPs) (haplotypes) (8). Given a focal population within a larger data set containing many such groups, the chromosomes of individuals in this population share ancestors with those in other populations, resulting in shared “chunks” of DNA. We used CHROMOPAINTER (8) to decompose each chromosome as a series of haplotypic chunks, each inferred to be shared with an individual from one of the other groups and colored (or painted) by this group (Fig. 1B). If the focal population is admixed, the changing colors along a chromosome noisily reflect true but unknown underlying ancestry (Fig. 1B) and so can be used to learn details of the source group(s) involved. To do this, we modeled haplotypes within each unsampled source group as being found across a weighted mixture of sampled “donor” populations (9). If a source group is genetically relatively similar to a single sampled population, then this population will dominate the inferred mixture. If there is no close proxy for the admixing group in the sample, especially likely for ancient admixture events or sparsely sampled regions, several donor populations will be needed to approximate its pattern of haplotype sharing. The focal population is then automatically a haplotype mixture of the combined donors, because it is a mixture of the source groups. Inferring the reduced set of groups within the mixture allows us to produce a “cleaned” painting (Fig. 1B) using only these groups.

To assess the evidence for admixture and date events, informally we measured the scale at which the cleaned painting changes along the genome. Specifically, we produced a coancestry curve for each pair of donor populations, plotting genetic distance x against a measure of how often a pair of haplotype chunks separated by x come from each respective population, analogously to ROLLOFF curves (4), and averaging over uncertain and typically computationally estimated haplotype phase (9). In theory, given a single admixture event, ancestry chunks inherited from each source have an exponential size distribution, resulting in an exponential decay of these coancestry curves (9). The rate of decay in all curves will be equal to the time in generations since admixture (Fig. 1C) (4, 9, 10), allowing estimation of this date: Steeper decay corresponds to older admixture. Such a decay distinguishes true admixture from ancient spatial structure and should only occur in recipient but not donor groups involved in nonreciprocal admixture events. We test for evidence to reject ($P < 0.01$) a no-admixture event (Fig. 1B).
null model, that is, no exponential decay in (normalized) coancestry curves, via bootstrapping (9). Multiple admixture times result in a mixture of exponentials (9); if admixture is detected, we test for evidence of multiple admixture times (e.g., two episodes of admixture or more continuous admixture over a longer period; empirical $P < 0.05$ in simulations), comparing the fit of a single exponential decay rate versus a mixture of rates.

The curve heights (intercepts) provide complementary information to deconvolve the number and genetic composition of the ancestral sources before admixture (11). Fitted curves for all pairs of donor groups (Fig. 1C shows three examples) specify a pairwise intercept matrix, which, after normalization, we decompose by using a series of eigenvectors. Analogous to the standard use of eigenvector decomposition in principal components analysis (PCA) in genetics to estimate relative ancestry source contributions.
for different individuals (12), the eigenvectors allow us to estimate the relative contribution to different admixing sources (e.g., source 1 versus source 2) for each different donor group (9). Also as for PCA, admixture between K distinct source populations produces K – 1 significant eigenvectors (13), and we test for three or more admixing sources by testing (empirically) for evidence of two or more such eigenvectors (P < 0.05) (9). After iterative modeling to improve results, this allows us to attempt to “reverse” the admixture process (Fig. 1D) and to infer the haplotypic makeup of admixing source groups as well as admixture date(s) in our method, which we call GLOBETROTTER.

Simulations
To test our approach under diverse single, complex, and no-admixture scenarios, incorporating many of the complexities (such as unsampled or admixed donor groups) likely to be present in real data, we simulated admixture scenarios involving real (but hidden to our analysis) human populations (4, 9) and populations generated under a coalescent framework (14) incorporating inferred (15–18) past demographic events. Admixture was simulated between 7 and 160 generations [200 to 4400 years, assuming 28 years per human generation (19)] ago, with admixture fractions 3 to 50% and genetic differentiation (FST) between the admixing groups varying from 0.018 (similar to Europe versus Central Asia) to 0.185 (similar to West Africa versus Europe). Results are detailed online (figs. S3 to S7 and tables S1 and S5). All populations simulated without admixture, including those with long-term migration, showed no admixture evidence (P > 0.1). Power to detect admixture (P < 0.01) when present was 94%, and 95% of our 95% bootstrap confidence intervals (CIs) contained the true admixture date, including cases with two distinct incidents of admixture or multiple groups admixing simultaneously. Inferred source accuracy was very high (9), with, for example, the mixture representation predicting a haplotype composition more correlated to the true, typically unsampled, source population than to any single sampled population >80% of the time. However, source accuracy was lower for admixing sources contributing only 5% of DNA, with around 40% of such scenarios yielding elevated (>25%) rates of falsely inferring multiple admixture times and/or admixing groups. Further testing demonstrated robustness of GLOBETROTTER, in simulations and real data, to haplotypic phase inference approach used, inclusion/exclusion of particular chromosomes, genetic map chosen to provide genetic distances, and the presence of population bottlenecks since admixture, whereas GLOBETROTTER admixture dating was improved relative to ROLLOFF (4, 9).

Nevertheless, there are multiple settings that we believe are challenging for our approach. First, although the admixing sources need not be sampled—often impossible because of genetic drift, extinction, or later admixture into the sources themselves—source inference is improved when more similar extant groups are sampled, and GLOBETROTTER may miss events where we lack any extant group that can separate sources. Second, sampling of several genetically very similar groups can mask admixture events they share. Similarly, a caveat is that where genuine, recent bidirectional gene flow has occurred, admixture fractions are difficult to define and interpret. However, date estimation is predicted to still be useful, and in real data the majority of our inferred events do not appear to be bidirectional in this manner. Third, even in theory our approach finds it challenging to distinguish distinct continuous “pulses” of admixture and continuous migration over some time frame (9), because of the difficulty of separating exponential mixtures (20). If the time frame were narrow, we expect to infer a single admixture time within the range of migration dates. Where we infer two admixture dates, in particular with the same source groups, the exponential decay signal could also be consistent with more continuous migration, and so we conservatively refer to this as admixture at multiple dates. Last, we only attempt to analyze populations with signals consistent with at most three groups admixing and infer at most two admixture times, and we can provide only less precise inference of sources for the weaker or older admixture signal in these complex cases (9).

Analysis of Worldwide Admixture
By using GLOBETROTTER, we analyzed 1490 individuals from 95 worldwide human groups (table S10 and fig. S12) (9), composed of 17 newly genotyped groups (21), 53 from the Human Genome Diversity Panel (HGDP) (22), and 25 from other sources (23, 24), filtered to 474,491 autosomal SNPs. We phased the individuals by using IMPUTE2 (9, 25) and used fineSTRUCTURE (8) to verify homogeneity within labeled populations, to identify genetically similar and clustered groups, and to remove outlying individuals (figs. S12 to S14 and tables S10 and S11). Of the 95 populations, 80 showed evidence (P < 0.01) of admixture, although nine could not be characterized by our approach (table S12). More than half of these have evidence of multiple waves of admixture (P < 0.05), and estimated admixture times vary from <10 to >150 generations (Fig. 2). We present individual results, for each population, via an interactive map online (26). We tested consistency of our results against a previous analysis of the 53 groups within the HGDP (11), which identified 34 groups with evidence of recent admixture. We identified (P < 0.01) admixture evidence in all 34 cases (with multiple event evidence in 15 cases) and obtained 95% admixture date CIs narrower than, but consistent with, those estimated by using ROLLOFF (9, 11). For 10 of 19 HGDP groups lacking previous support for recent admixture, GLOBETROTTER also identifies no events: In the remaining populations, admixture is inferred as occurring between genetically similar sources (FST < 0.02), a challenging setting where simulations suggest our method is more powerful (9).

In several instances, GLOBETROTTER clarifies or extends previous genetic analyses. For example, a previous study (27) inferred admixture in the Maya, with best source populations the Mozabites from North Africa and the Native American Suri, speculating on the basis of historical events that this might actually represent a mixture of European, West African, and Native American ancestry sources. GLOBETROTTER inferred admixture between three groups in the Maya dating to around 1670 CE (9 generations ago) (28) (Fig. 2, A and D, fuchsia box 1), with distinct sources from Europe (most genetically similar to the Spanish), West Africa (the Yoruba), and the Americas (the Pima, the nearest sampled group in the Americas). A different method, which aims to detect but not date admixture, concluded that Cambodians trace ~16% of their DNA to a group equally related to modern-day Europeans and East Asians (29). GLOBETROTTER infers a ~19% contribution from a similar source related to modern-day Central, South, and East Asians and an ~81% contribution from a source related specifically to modern-day Han and Dai, the latter a branch of the Tai people who entered the region in historical times (30) (Fig. 2D, orange box 5). Further, this event dates to 1362 CE (1194 to 1502 CE), a period spanning the end of the Indianized Khmer empire (802 to 1431 CE) (30), one of the most powerful empires in Southeast Asia, whose fall was hypothesized to relate to a Tai influx (30).

A comparison with the historical record becomes progressively more difficult for older episodes. Even when events are well attested, their exact genetic impacts (if any) are rarely if ever known, motivating our approach. Nevertheless, we have identified nine groups of populations showing related events, incorporating almost all (19/20) with the strongest GLOBETROTTER admixture evidence (9). Results are presented as online maps (26). Some events appear to match well with particular historical occurrences, such as the so-called Bantu Expansion into Southern Africa (9). Events affecting a group of seven populations (Fig. 2D, purple box 4) correspond in time to the rapid expansion, initiated by Genghis Khan, of the Mongol empire (1206 to 1368 CE) (31), one of the most dramatic events in human history. These populations, including the Hazara (32, 33), the Uygur (34), and the Mongola themselves, were sampled from within the range of the Mongol empire and show an admixture event dating within the Mongol Period, with one source closely genetically related to the Mongola that progressively decreases in proportion westward, to 8% in the Turkish (Fig. 2D).

Seventeen populations from the Mediterranean, the Near East, and countries bordering the Arabian Sea (Fig. 2D, blue box 3) show signals of admixture from sub-Saharan Africa, with most recent dates in the range 890 to 1754 CE (Fig. 2, B
and D). We interpret these signals, consistent with overlapping results of previous studies (4, 20), as resulting from the Arab expansion and slave trade, which originated around the seventh century CE (35). Our event dates are highly consistent with this but also imply earlier sub-Saharan African gene flow into, for example, the Moroccans. The highest-contributing sub-Saharan donor is West African for all 12 Mediterranean populations and an East or South African Bantu-speaking group for all five Arabian Sea populations (Fig. 2D), confirming genetically different sources for these slave trades (35).

A population group centered around Eastern Europe shows signals of complex admixture. FineSTRUCTURE did not fully separate groups from this region, suggesting masked shared events might be present. We therefore reinterpreted them excluding each other as donors: We performed similar reanalyses of five additional geographic regions for the same reason (table S16 and figs. S16 to S21). The easternly Russians and Chuvash both show evidence (P < 0.05) of admixture at more than one time (Fig. 2D), at least partially predating the Mongol empire, between groups with ancestry related to Northeast Asians (e.g., the Oroqen, Mongola, and Yukut) and Europeans, respectively (table S16). Six other European populations (Fig. 2D, pink/maroon box 2) independently show evidence after the repainting for similar admixture events involving more than two groups (P < 0.02) at about the same time (Fig. 3). CIs for the admixture time(s) overlap but predate the Mongol empire, with estimates from 440 to 1080 CE (Fig. 3). In each population, one source group has at least some ancestry related to Northeast Asians (e.g., the Oroqen, Mongola, and Yukut) and Europeans, respectively (table S16). Six other European populations (Fig. 2D, pink/maroon box 2) independently show evidence after the repainting for similar admixture events involving more than two groups (P < 0.02) at about the same time (Fig. 3). CIs for the admixture time(s) overlap but predate the Mongol empire, with estimates from 440 to 1080 CE (Fig. 3).

Within the past several thousand years affect most human populations, and this needs to be taken into account in inferences aiming to look at the more distant history of our species. Future improvements in whole-genome sequencing, greater sample sizes, and incorporation of ancient DNA, together with additional methodological extensions, are likely to allow better understanding of ancient events where little or no historical record exists, to identify many additional events, to infer sex biases, and to provide more precise event characterization than currently possible. We believe our approach will extend naturally to these settings, as well as to other species.

References and Notes
9. Information on materials and methods is available on Science Online.
27. N. Patterson et al., Genetics 192, 1065–1093 (2012).
28. We converted g inferred generations to the admixture year: 1590 = (g + 1) × 28.

Acknowledgments: We are grateful for the John Fell Fund—University of Oxford, the NIH, the Wellcome Trust (S.M., grant 098387/Z/12/Z), the Biotechnology and Biological Sciences Research Council, the Royal Society/Wellcome Trust (G.H., grant 098386/Z/12/Z), and the Istituto Italiano di Antropologia for funding. J.F.W. is a director, stockholder, and employee of ScotlandDNA (and formerly of EthnoAncestry). We thank S. Karakanach, D. Toncheva, P. Anagnostou, F. Calli, F. Brighelli, V. Romano, G. LeFranc, C. Buresi, J. Ben Chibani, A. Huj-Kheil, S. Denden, R. Ploski, T. Herzig, T. Moen, P. Krajewski, and R. Herrera for providing samples for our genotyping and the blood donors and the staff of the Unità Operativa Complessa di Medicina Trafuscinale, Azienda Ospedaliero Umberta I, Siracusa (Italy). Data analyzed in this study may be downloaded via http://admixturemap.
41. Raw genotype data are available at the Gene Expression Omnibus database online (www.ncbi.nlm.nih.gov/geo/), series accession number GSE53626.
42. Supplementary Materials
43. www.sciencemag.org/content/343/6172/2741/DC1
44. Materials and Methods
45. Supplementary Text
46. Figs. S1 to S21
47. Tables S1 to S16
48. Appendix
49. References (41–82)
50. 22 July 2013; accepted 20 December 2013 10.1126/science.1243518