Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow

Mark Lipson,[†,1] Po-Ru Loh,[†,1] Alex Levin,[1] David Reich,[2,3] Nick Patterson,[2] and Bonnie Berger[*,1,2]

[1]Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
[2]Broad Institute, Cambridge, Massachusetts
[3]Department of Genetics, Harvard Medical School
[†]These authors contributed equally to this work.
*Corresponding author: E-mail: bab@mit.edu.
Associate editor: John Novembre

Abstract

The recent explosion in available genetic data has led to significant advances in understanding the demographic histories of and relationships among human populations. It is still a challenge, however, to infer reliable parameter values for complicated models involving many populations. Here, we present *MixMapper*, an efficient, interactive method for constructing phylogenetic trees including admixture events using single nucleotide polymorphism (SNP) genotype data. *MixMapper* implements a novel two-phase approach to admixture inference using moment statistics, first building an unadmixed scaffold tree and then adding admixed populations by solving systems of equations that express allele frequency divergences in terms of mixture parameters. Importantly, all features of the model, including topology, sources of gene flow, branch lengths, and mixture proportions, are optimized automatically from the data and include estimates of statistical uncertainty. *MixMapper* also uses a new method to express branch lengths in easily interpretable drift units. We apply *MixMapper* to recently published data for Human Genome Diversity Cell Line Panel individuals genotyped on a SNP array designed especially for use in population genetics studies, obtaining confident results for 30 populations, 20 of them admixed. Notably, we confirm a signal of ancient admixture in European populations— including previously undetected admixture in Sardinians and Basques—involving a proportion of 20–40% ancient northern Eurasian ancestry.

Key words: admixture, human populations, genetic drift, moment statistics.

Introduction

The most basic way to represent the evolutionary history of a set of species or populations is through a phylogenetic tree, a model that in its strict sense assumes that there is no gene flow between populations after they have diverged (Cavalli-Sforza and Edwards 1967). In many settings, however, groups that have split from one another can still exchange genetic material. This is certainly the case for human population history, during the course of which populations have often diverged only incompletely or diverged and subsequently mixed again (Reich et al. 2009; Wall et al. 2009; Green et al. 2010; Laval et al. 2010; Reich et al. 2010; Gravel et al. 2011; Patterson et al. 2012). To capture these more complicated relationships, previous studies have considered models allowing for continuous migration among populations (Wall et al. 2009; Laval et al. 2010; Gravel et al. 2011) or have extended simple phylogenetic trees into *admixture trees*, in which populations on separate branches are allowed to remerge and form an admixed offspring population (Chikhi et al. 2001; Wang 2003; Reich et al. 2009; Sousa et al. 2009; Patterson et al. 2012). Both of these frameworks, of course, still represent substantial simplifications of true population histories, but they can help capture a range of new and interesting phenomena.

Several approaches have previously been used to build phylogenetic trees incorporating admixture events from genetic data. First, likelihood methods (Chikhi et al. 2001; Wang 2003; Sousa et al. 2009) use a full probabilistic evolutionary model, which allows a high level of precision with the disadvantage of greatly increased computational cost. Consequently, likelihood methods can in practice only accommodate a small number of populations (Wall et al. 2009; Laval et al. 2010; Gravel et al. 2011; Sirén et al. 2011). Moreover, the tree topology must generally be specified in advance, meaning that only parameter values can be inferred automatically and not the arrangement of populations in the tree. By contrast, the moment-based methods of Reich et al. (2009) and Patterson et al. (2012) use only means and variances of allele frequency divergences. Moments are simpler conceptually and especially computationally, and they allow for more flexibility in model conditions. Their disadvantages can include reduced statistical power and difficulties in designing precise estimators with desirable statistical properties (e.g., unbiasedness) (Wang 2003). Finally, a number of studies have considered "phylogenetic networks," which generalize trees to include cycles and multiple edges between pairs of nodes and can be used to model population histories involving hybridization (Huson and Bryant 2006; Yu et al. 2012).

Open Access

Article

Fast Track

However, these methods also tend to be computationally expensive.

In this work, we introduce *MixMapper*, a new computational tool that fits admixture trees by solving systems of moment equations involving the pairwise distance statistic f_2 (Reich et al. 2009; Patterson et al. 2012), which is the average squared allele frequency difference between two populations. The theoretical expectation of f_2 can be calculated in terms of branch lengths and mixture fractions of an admixture tree and then compared with empirical data. *MixMapper* can be thought of as a generalization of the *qpgraph* package (Patterson et al. 2012), which takes as input genotype data, along with a proposed arrangement of admixed and unadmixed populations, and returns branch lengths and mixture fractions that produce the best fit to allele frequency moment statistics measured on the data. *MixMapper*, by contrast, performs the fitting in two stages, first constructing an unadmixed scaffold tree via neighbor-joining and then automatically optimizing the placement of admixed populations onto this initial tree. Thus, no topological relationships among populations need to be specified in advance.

Our method is similar in spirit to the independently developed *TreeMix* package (Pickrell and Pritchard 2012). Like *MixMapper*, *TreeMix* builds admixture trees from second moments of allele frequency divergences, although it does so via a composite likelihood maximization approach made tractable with a multivariate normal approximation. Procedurally, *TreeMix* initially fits a full set of populations as an unadmixed tree, and gene flow edges are added sequentially to account for the greatest errors in the fit (Pickrell and Pritchard 2012). This format makes *TreeMix* well suited to handling very large trees: the entire fitting process is automated and can include arbitrarily many admixture events simultaneously. In contrast, *MixMapper* begins with a carefully screened unadmixed scaffold tree to which admixed populations are added with best-fitting parameter values, an interactive design that enables precise modeling of particular populations of interest.

We use *MixMapper* to model the ancestral relationships among 52 populations from the CEPH-Human Genome Diversity Cell Line Panel (HGDP) (Rosenberg et al. 2002; Li et al. 2008) using recently published data from a new, specially ascertained single nucleotide polymorphism (SNP) array designed for population genetics applications (Keinan et al. 2007; Patterson et al. 2012). Previous studies of these populations have built simple phylogenetic trees (Li et al. 2008; Sirén et al. 2011), identified a substantial number of admixed populations with likely ancestors (Patterson et al. 2012), and constructed a large-scale admixture tree (Pickrell and Pritchard 2012). Here, we add an additional level of quantitative detail, obtaining best-fit admixture parameters with bootstrap error estimates for 30 HGDP populations, of which 20 are admixed. The results include, most notably, a significant admixture event (Patterson et al. 2012) in the history of all sampled European populations, among them Sardinians and Basques.

New Approaches

The central problem we consider is as follows: given an array of SNP data sampled from a set of individuals grouped by population, what can we infer about the admixture histories of these populations using simple statistics that are functions of their allele frequencies? Methodologically, the *MixMapper* workflow (fig. 1) proceeds as follows. We begin by computing f_2 distances between all pairs of study populations, from which we construct an unadmixed phylogenetic subtree to serve as a scaffold for subsequent mixture fitting. The choice of populations for the scaffold is done via initial filtering of populations that are clearly admixed according to the three-population test (Reich et al. 2009; Patterson et al. 2012), followed by selection of a subtree that is approximately additive along its branches, as is expected in the absence of admixture (Materials and Methods and supplementary text S1, Supplementary Material online, for full details).

Next, we expand the model to incorporate admixtures by attempting to fit each population not in the scaffold as a mixture between some pair of branches of the scaffold. Putative admixtures imply algebraic relations among f_2 statistics, which we test for consistency with the data, allowing us to identify likely sources of gene flow and estimate mixture parameters (fig. 2; supplementary text S1, Supplementary Material online). After determining likely two-way admixture events, we further attempt to fit remaining populations as
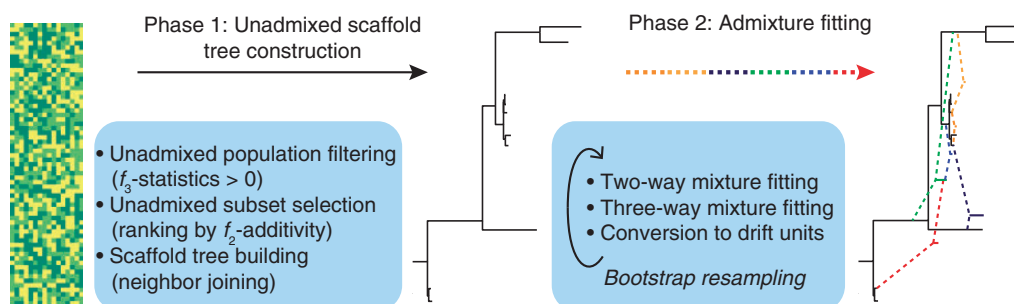
Fig. 1. *MixMapper* workflow. *MixMapper* takes as input an array of SNP calls annotated with the population to which each individual belongs. The method then proceeds in two phases, first building a tree of (approximately) unadmixed populations and then attempting to fit the remaining populations as admixtures. In the first phase, *MixMapper* produces a ranking of possible unadmixed trees in order of deviation from f_2-additivity; based on this list, the user selects a tree to use as a scaffold. In the second phase, *MixMapper* tries to fit remaining populations as two- or three-way mixtures between branches of the unadmixed tree. In each case, *MixMapper* produces an ensemble of predictions via bootstrap resampling, enabling confidence estimation for inferred results.
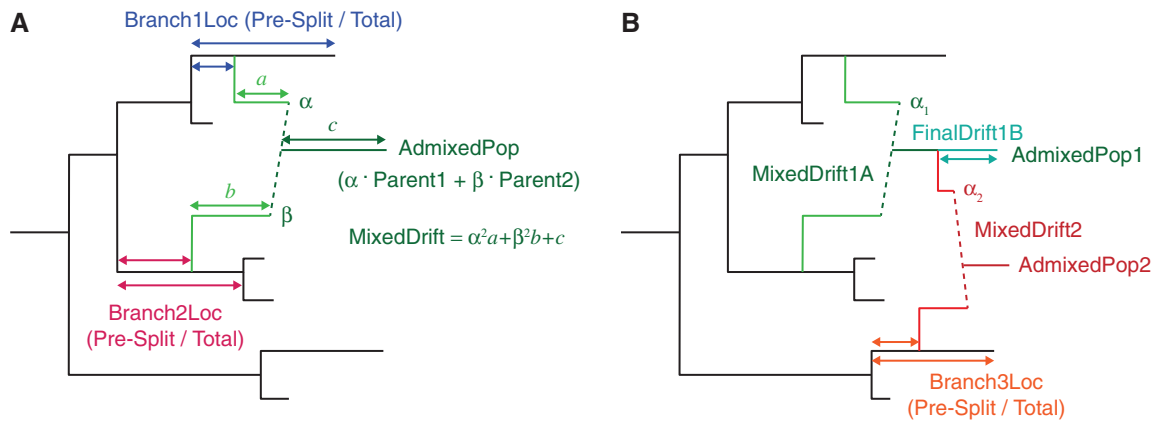
Fig. 2. Schematic of mixture parameters fit by *MixMapper*. (*A*) A simple two-way admixture. *MixMapper* infers four parameters when fitting a given population as an admixture. It finds the optimal pair of branches between which to place the admixture and reports the following: Branch1Loc and Branch2Loc are the points at which the mixing populations split from these branches (given as pre-split length/total branch length); α is the proportion of ancestry from Branch1 ($\beta = 1 - \alpha$ is the proportion from Branch2); and MixedDrift is the linear combination of drift lengths $\alpha^2 a + \beta^2 b + c$. (*B*) A three-way mixture: here AdmixedPop2 is modeled as an admixture between AdmixedPop1 and Branch3. There are now four additional parameters; three are analogous to the above, namely, Branch3Loc, α_2, and MixedDrift2. The remaining degree of freedom is the position of the split along the AdmixedPop1 branch, which divides MixedDrift into MixedDrift1A and FinalDrift1B.

three-way mixtures involving the inferred two-way mixed populations, by similar means. Finally, we use a new formula to convert the f_2 tree distances into absolute drift units (supplementary text S2, Supplementary Material online). Importantly, we apply a bootstrap resampling scheme (Efron 1979; Efron and Tibshirani 1986) to obtain ensembles of predictions, rather than single values, for all model variables. This procedure allows us to determine confidence intervals for parameter estimates and guard against overfitting. For a data set on the scale of the HGDP, after initial setup time on the order of an hour, *MixMapper* determines the best-fit admixture model for a chosen population in a few seconds, enabling real-time interactive investigation.
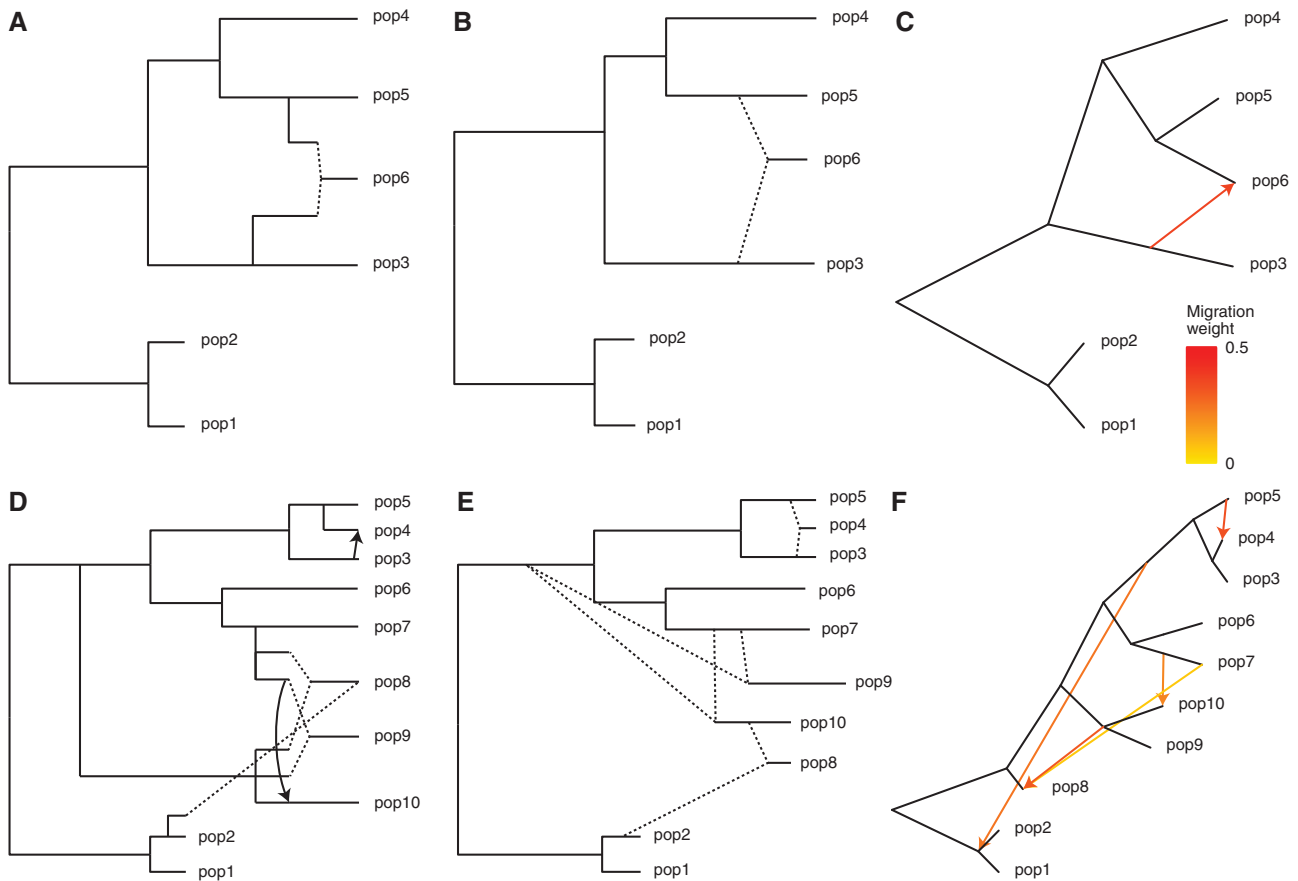
Results

Simulations

To test the inference capabilities of *MixMapper* on populations with known histories, we ran it on two data sets generated with the coalescent simulator ms (Hudson 2002) and designed to have similar parameters to our human data. In both cases, we simulated 500 regions of 500 kb each for 25 diploid individuals per population, with an effective population size of 5,000 or 10,000 per population, a mutation rate of 0.5×10^{-8} per base per generation (intentionally low so as not to create unreasonably many SNPs), and a recombination rate of 10^{-8} per base per generation. Full ms commands can be found in Materials and Methods. We ascertained SNPs present at minor allele frequency 0.05 or greater in an outgroup population and then removed that population from the analysis.

For the first admixture tree, we simulated six nonoutgroup populations, with one of them, pop6, admixed (fig. 3A). Applying *MixMapper*, no admixtures were detected with the three-population test, but the most additive subset with at least five populations excluded pop6 (max deviation from additivity 2.0×10^{-4} vs. second-best 7.7×10^{-4}; see

Materials and Methods), so we used this subset as the scaffold tree. We then fit pop6 as admixed, and *MixMapper* recovered the correct gene flow topology with 100% confidence and inferred the other parameters of the model quite accurately (fig. 3B; supplementary table S1, Supplementary Material online). For comparison, we also analyzed the same data with *TreeMix* and again obtained accurate results (fig. 3C).

For the second test, we simulated a complex admixture scenario involving 10 nonoutgroup populations, with six unadmixed and four admixed (fig. 3D). In this example, pop4 is recently admixed between pop3 and pop5, but over a continuous period of 40 generations. Meanwhile, pop8, pop9, and pop10 are all descended from older admixture events, which are similar but with small variations (lower mixture fraction in pop9, 40-generation continuous gene flow in pop10, and subsequent pop2-related admixture into pop8). In the first phase of *MixMapper*, the recently admixed pop4 and pop8 were detected with the three-population test. From among the other eight populations, a scaffold tree consisting of pop1, pop2, pop3, pop5, pop6, and pop7 provided thorough coverage of the data set and was more additive (max deviation 3.5×10^{-4}) than the second-best six-population scaffold (5.4×10^{-4}) and the best seven-population scaffold (1.2×10^{-3}). Using this scaffold, *MixMapper* returned very accurate and high-confidence fits for the remaining populations (fig. 3E; supplementary table S1, Supplementary Material online), with the correct gene flow topologies inferred with 100% confidence for pop4 and pop10, 98% confidence for pop9, and 61% confidence for pop8 (fit as a three-way admixture; 39% of replicates placed the third gene flow source on the branch adjacent to pop2, as shown in supplementary table S1, Supplementary Material online). In contrast, *TreeMix* inferred a less accurate admixture model for this data set (fig. 3F). *TreeMix* correctly identified pop4 as admixed, and it placed three migration edges among pop7, pop8, pop9, and pop10, but two of the

FIG. 3. Results with simulated data. (A–C) First simulated admixture tree, with one admixed population. Shown are (A) the true phylogeny, (B) *MixMapper* results, and (C) *TreeMix* results. (D–F) Second simulated admixture tree, with four admixed populations. Shown are (D) the true phylogeny, (E) *MixMapper* results, and (F) *TreeMix* results. In (A) and (D), dotted lines indicate instantaneous admixtures, whereas arrows denote continuous (unidirectional) gene flow over 40 generations. Both *MixMapper* and *TreeMix* infer point admixtures, depicted with dotted lines in (B) and (E) and colored arrows in (C) and (F). In (B) and (E), the terminal drift edges shown for admixed populations represent half the total mixed drift. Full inferred parameters from *MixMapper* are given in supplementary table S1, Supplementary Material online.

five total admixtures (those originating from the common ancestor of pops 3–5 and the common ancestor of pops 9–10) did not correspond to true events. Also, *TreeMix* did not detect the presence of admixture in pop9 or the pop2-related admixture in pop8.

Application of *MixMapper* to HGDP Data

Despite the focus of the HGDP on isolated populations, most of its 53 groups exhibit signs of admixture detectable by the three-population test, as has been noted previously (Patterson et al. 2012). Thus, we hypothesized that applying *MixMapper* to this data set would yield significant insights. Ultimately, we were able to obtain comprehensive results for 20 admixed HGDP populations (fig. 4), discussed in detail in the following sections.

Selection of a 10-Population Unadmixed Scaffold Tree

To construct an unadmixed scaffold tree for the HGDP data to use in fitting admixtures, we initially filtered the list of 52 populations (having removed San due to ascertainment of our SNP panel in a San individual; see Materials and Methods) with the three-population test, leaving only 20 that are potentially unadmixed. We further excluded Mbuti and Biaka

Pygmies, Kalash, Melanesian, and Colombian from the list of candidate populations due to external evidence of admixture (Loh et al. 2013).

It is desirable to include a wide range of populations in the unadmixed scaffold tree to provide both geographic coverage and additional constraints that facilitate the fitting of admixed populations (see Materials and Methods). Additionally, incorporating at least four continental groups provides a fairer evaluation of additivity, which is roughly equivalent to measuring discrepancies in fitting phylogenies to quartets of populations. If all populations fall into three or fewer tight clades, however, any quartet must contain at least two populations that are closely related. At the same time, including too many populations can compromise the accuracy of the scaffold. We required that our scaffold tree include representatives of at least four of the five major continental groups in the HGDP data set (Africa, Europe, Oceania, Asia, and the Americas), with at least two populations per group (when available) to clarify the placement of admixing populations and improve the geographical balance. Subject to these conditions, we selected an approximately unadmixed scaffold tree containing 10 populations, which we found to provide a good balance between additivity and
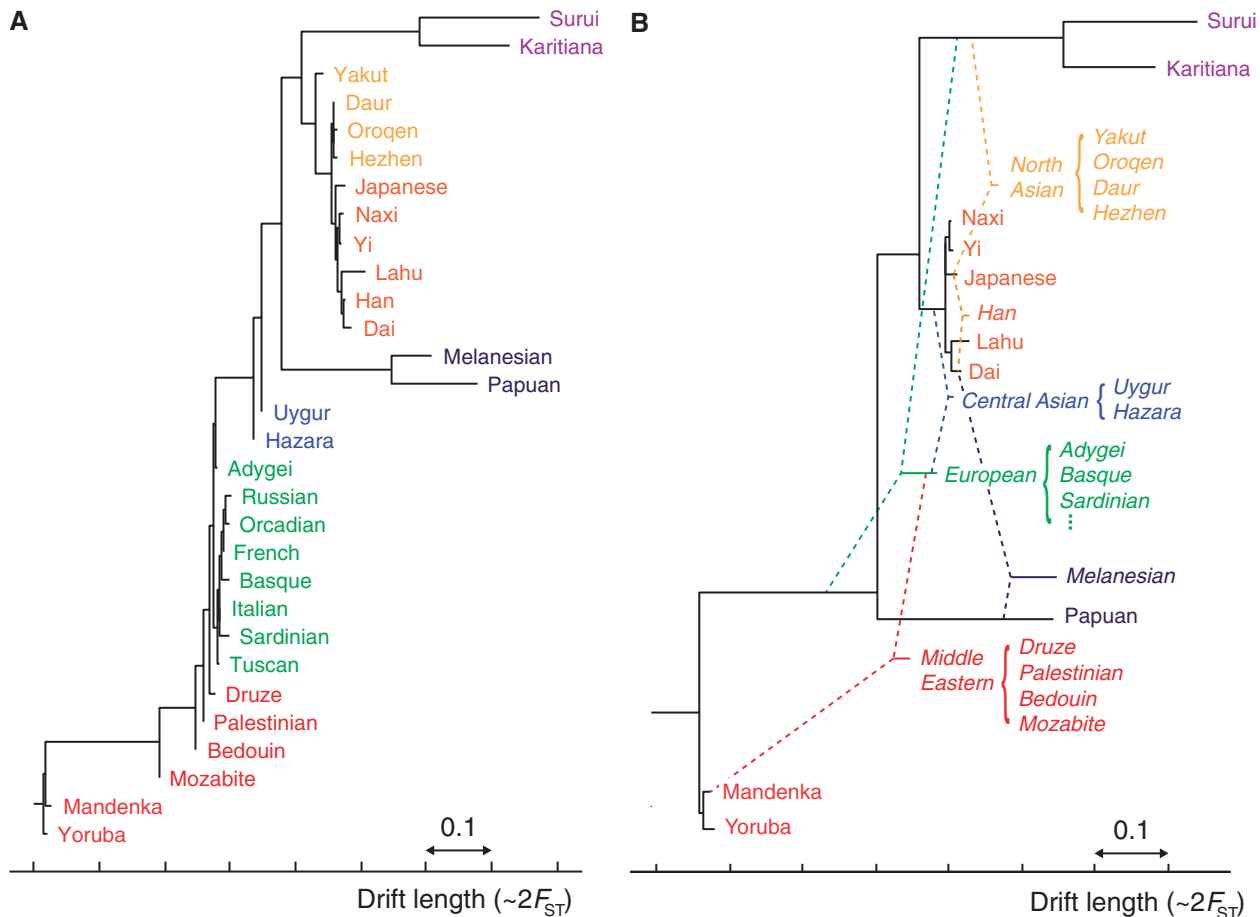
Fig. 4. Aggregate phylogenetic trees of HGDP populations with and without admixture. (A) A simple neighbor-joining tree on the 30 populations for which *MixMapper* produced high-confidence results. This tree is analogous to the one given by (Li et al. 2008, fig. 1B), and the topology is very similar. (B) Results from *MixMapper*. The populations appear in roughly the same order, but the majority are inferred to be admixed, as represented by dashed lines (cf. Pickrell and Pritchard 2012 and supplementary fig. S4, Supplementary Material online). Note that drift units are not additive, so branch lengths should be interpreted individually.

comprehensiveness: Yoruba, Mandenka, Papuan, Dai, Lahu, Japanese, Yi, Naxi, Karitiana, and Suruí (fig. 4B). These populations constitute the second-most additive (max deviation 1.12×10^{-3}) of 21 similar trees differing only in which East Asian populations are included (range $1.12–1.23 \times 10^{-3}$); we chose them over the most-additive tree because they provide slightly better coverage of Asia. To confirm that modeling these 10 populations as unadmixed in *MixMapper* is sensible, we checked that none of them can be fit in a reasonable way as an admixture on a tree built with the other nine (Materials and Methods). Furthermore, we repeated all of the analyses to follow using nine-population subsets of the unadmixed tree as well as an alternative 11-population tree and confirmed that our results are robust to the choice of scaffold (supplementary figs. S1 and S2, and tables S2–S4, Supplementary Material online).

Ancient Admixture in the History of Present-Day European Populations

A notable feature of our unadmixed scaffold tree is that it does not contain any European populations. Patterson et al. (2012) previously observed negative f_3 values indicating admixture in all HGDP Europeans other than Sardinian and

Basque. Our *MixMapper* analysis uncovered the additional observation that potential trees containing Sardinian or Basque along with representatives of at least three other continents are noticeably less additive than four-continent trees of the same size without Europeans: from our set of 15 potentially unadmixed populations, none of the 100 most additive 10-population subtrees include Europeans. This points to the presence of admixture in Sardinian and Basque as well as the other European populations.

Using *MixMapper*, we added European populations to the unadmixed scaffold via admixtures (fig. 5 and table 1). For all eight groups in the HGDP data set, the best fit was as a mixture of a population related to the common ancestor of Karitiana and Suruí (in varying proportions of ~20–40%, with Sardinian and Basque among the lowest and Russian the highest) with a population related to the common ancestor of all non-African populations on the tree. We fit all eight European populations independently, but notably, their ancestors branch from the scaffold tree at very similar points, suggesting a similar broad-scale history. Their branch positions are also qualitatively consistent with previous work that used the 3-population test to deduce ancient admixture for Europeans other than Sardinian and Basque

A

B
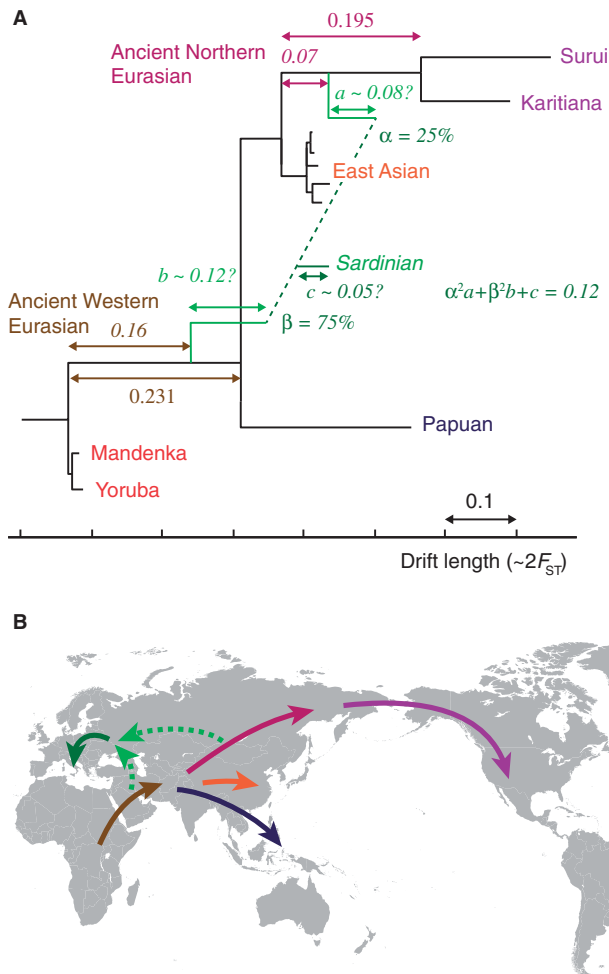
FIG. 5. Inferred ancient admixture in Europe. (A) Detail of the inferred ancestral admixture for Sardinians (other European populations are similar). One mixing population splits from the unadmixed tree along the common ancestral branch of Native Americans ("Ancient Northern Eurasian") and the other along the common ancestral branch of all non-Africans ("Ancient Western Eurasian"). Median parameter values are shown; 95% bootstrap confidence intervals can be found in table 1. The branch lengths *a*, *b*, and *c* are confounded, so we show a plausible combination. (B) Map showing a sketch of possible directions of movement of ancestral populations. Colored arrows correspond to labeled branches in (A).

(Patterson et al. 2012). To confirm the signal in Sardinian and Basque, we applied f_4 ratio estimation (Reich et al. 2009; Patterson et al. 2012), which uses allele frequency statistics in a simpler framework to infer mixture proportions. We estimated approximately 20–25% "ancient northern Eurasian" ancestry (supplementary table S5, Supplementary Material online), which is in very good agreement with our findings from *MixMapper* (table 1).

At first glance, this inferred admixture might appear improbable on geographical and chronological grounds, but importantly, the two ancestral branch positions do not represent the mixing populations themselves. Rather, there may be substantial drift from the best-fit branch points to the true mixing populations, indicated as branch lengths *a* and *b* in figure 5A. Unfortunately, these lengths, along with the postadmixture drift *c*, appear only in a fixed linear combination in the system of f_2 equations (supplementary text S1, Supplementary Material online), and current methods can only give estimates of this linear combination rather than the individual values (Patterson et al. 2012). One plausible arrangement, however, is shown in figure 5A for the case of Sardinian.

Two-Way Admixtures Outside of Europe

We also found several other populations that fit robustly onto the unadmixed tree using simple two-way admixtures (table 2). All of these can be identified as admixed using the 3-population or 4-population tests (Patterson et al. 2012), but with *MixMapper*, we are able to provide the full set of best-fit parameter values to model them in an admixture tree.

First, we found that four populations from North-Central and Northeast Asia—Daur, Hezhen, Oroqen, and Yakut—are likely descended from admixtures between native North Asian populations and East Asian populations related to Japanese. The first three are estimated to have roughly 10–30% North Asian ancestry, whereas Yakut has 50–75%. Melanesians fit optimally as a mixture of a Papuan-related population with an East Asian population close to Dai, in a proportion of roughly 80% Papuan-related, similar to previous estimates (Reich et al. 2011; Xu et al. 2012). Finally, we

Table 1. Mixture Parameters for Europeans.

AdmixedPop	No. of Replicates[a]	α[b]	Branch1Loc (Anc. N. Eurasian)[c]	Branch2Loc (Anc. W. Eurasian)[c]	MixedDrift[d]
Adygei	500	0.254–0.461	0.033–0.078/0.195	0.140–0.174/0.231	0.077–0.092
Basque	464	0.160–0.385	0.053–0.143/0.196	0.149–0.180/0.231	0.105–0.121
French	491	0.184–0.386	0.054–0.130/0.195	0.149–0.177/0.231	0.089–0.104
Italian	497	0.210–0.415	0.043–0.108/0.195	0.137–0.173/0.231	0.092–0.109
Orcadian	442	0.156–0.350	0.068–0.164/0.195	0.161–0.185/0.231	0.096–0.113
Russian	500	0.278–0.486	0.045–0.091/0.195	0.146–0.181/0.231	0.079–0.095
Sardinian	480	0.150–0.350	0.045–0.121/0.195	0.146–0.176/0.231	0.107–0.123
Tuscan	489	0.179–0.431	0.039–0.118/0.195	0.137–0.177/0.231	0.088–0.110

[a]Number of bootstrap replicates (out of 500) placing the mixture between the two branches shown.
[b]Proportion of ancestry from "ancient northern Eurasian" (95% bootstrap confidence interval).
[c]See figure 5A for the definition of the "ancient northern Eurasian" and "ancient western Eurasian" branches in the scaffold tree; Branch1Loc and Branch2Loc are the points at which the mixing populations split from these branches (expressed as confidence interval for split point/branch total, as in figure 2A).
[d]Sum of drift lengths $\alpha^2 a + (1 - \alpha)^2 b + c$; see figure 2A.

Table 2. Mixture Parameters for Non-European Populations Modeled as Two-Way Admixtures.

AdmixedPop	Branch1 + Branch2[a]	No. of Replicates[b]	α[c]	Branch1Loc[d]	Branch2Loc[d]	MixedDrift[e]
Daur	Anc. N. Eurasian + Japanese	350	0.067–0.276	0.008–0.126/0.195	0.006–0.013/0.016	0.006–0.015
	Suruí + Japanese	112	0.021–0.058	0.008–0.177/0.177	0.005–0.010/0.015	0.005–0.016
Hezhen	Anc. N. Eurasian + Japanese	411	0.068–0.273	0.006–0.113/0.195	0.006–0.013/0.016	0.005–0.029
Oroqen	Anc. N. Eurasian + Japanese	410	0.093–0.333	0.017–0.133/0.195	0.005–0.013/0.015	0.011–0.030
	Karitiana + Japanese	53	0.025–0.086	0.014–0.136/0.136	0.004–0.008/0.016	0.008–0.026
Yakut	Anc. N. Eurasian + Japanese	481	0.494–0.769	0.005–0.026/0.195	0.012–0.016/0.016	0.030–0.041
Melanesian	Dai + Papuan	424	0.160–0.260	0.008–0.014/0.014	0.165–0.201/0.247	0.089–0.114
	Lahu + Papuan	54	0.155–0.255	0.003–0.032/0.032	0.167–0.208/0.249	0.081–0.114
Han	Dai + Japanese	440	0.349–0.690	0.004–0.014/0.014	0.008–0.016/0.016	0.002–0.006

[a]Optimal split points for mixing populations.
[b]Number of bootstrap replicates (out of 500) placing the mixture between Branch1 and Branch2; topologies are shown that that occur for at least 50 of 500 replicates.
[c]Proportion of ancestry from Branch1 (95% bootstrap confidence interval).
[d]Points at which mixing populations split from their branches (expressed as confidence interval for split point/branch total, as in figure 2A).
[e]Sum of drift lengths $\alpha^2 a + (1 - \alpha)^2 b + c$; see figure 2A.

Table 3. Mixture Parameters for Populations Modeled as Three-Way Admixtures.

AdmixedPop2	Branch3[a]	No. of Replicates[b]	α_2[c]	Branch3Loc[d]	MixedDrift1A[e]	FinalDrift1B[e]	MixedDrift2[e]
Druze	Mandenka	330	0.963–0.988	0.000–0.009/0.009	0.081–0.099	0.022–0.030	0.004–0.013
	Yoruba	82	0.965–0.991	0.000–0.010/0.010	0.080–0.099	0.022–0.029	0.005–0.013
	Anc. W. Eurasian	79	0.881–0.966	0.041–0.158/0.232	0.092–0.118	0.000–0.024	0.010–0.031
Palestinian	Anc. W. Eurasian	294	0.818–0.901	0.031–0.104/0.231	0.093–0.123	0.000–0.021	0.007–0.022
	Mandenka	146	0.909–0.937	0.000–0.009/0.009	0.083–0.097	0.022–0.029	0.001–0.007
	Yoruba	53	0.911–0.938	0.000–0.010/0.010	0.077–0.098	0.021–0.029	0.001–0.008
Bedouin	Anc. W. Eurasian	271	0.767–0.873	0.019–0.086/0.231	0.094–0.122	0.000–0.022	0.012–0.031
	Mandenka	176	0.856–0.923	0.000–0.008/0.008	0.080–0.099	0.023–0.030	0.006–0.018
Mozabite	Mandenka	254	0.686–0.775	0.000–0.009/0.009	0.088–0.109	0.012–0.022	0.017–0.032
	Anc. W. Eurasian	142	0.608–0.722	0.002–0.026/0.232	0.103–0.122	0.000–0.011	0.018–0.035
	Yoruba	73	0.669–0.767	0.000–0.008/0.010	0.086–0.108	0.012–0.023	0.017–0.031
Hazara	Anc. East Asian[f]	497	0.364–0.471	0.010–0.024/0.034	0.080–0.115	0.004–0.034	0.004–0.013
Uygur	Anc. East Asian[f]	500	0.318–0.438	0.007–0.023/0.034	0.088–0.123	0.000–0.027	0.000–0.009

[a]Optimal split point for the third ancestry component. The first two components are represented by a parent population splitting from the (admixed) Sardinian branch.
[b]Number of bootstrap replicates placing the third ancestry component on Branch3; topologies are shown that that occur for at least 50 of 500 replicates.
[c]Proportion of European-related ancestry (95% bootstrap confidence interval).
[d]Point at which mixing population splits from Branch3 (expressed as confidence interval for split point/branch total, as in figure 2A).
[e]Terminal drift parameters; see figure 2B.
[f]Common ancestral branch of the five East Asian populations in the unadmixed tree (Dai, Japanese, Lahu, Naxi, and Yi).

found that Han Chinese have an optimal placement as an approximately equal mixture of two ancestral East Asian populations, one related to modern Dai (likely more southerly) and one related to modern Japanese (likely more northerly), corroborating a previous finding of admixture in Han populations between northern and southern clusters in a large-scale genetic analysis of East Asia (HUGO Pan-Asian SNP Consortium 2009).

Recent Three-Way Admixtures Involving Western Eurasians

Finally, we inferred the branch positions of several populations that are well known to be recently admixed (cf. Patterson et al. 2012; Pickrell and Pritchard 2012) but for which one ancestral mixing population was itself anciently admixed in a similar way to Europeans. To do so, we applied

the capability of MixMapper to fit three-way admixtures (fig. 2B), using the anciently admixed branch leading to Sardinian as one ancestral source branch. First, we found that Mozabite, Bedouin, Palestinian, and Druze, in decreasing order of African ancestry, are all optimally represented as a mixture between an African population and an admixed western Eurasian population (not necessarily European) related to Sardinian (table 3). We also obtained good fits for Uygur and Hazara as mixtures between a western Eurasian population and a population related to the common ancestor of all East Asians on the tree (table 3).

Estimation of Ancestral Heterozygosity

Using SNPs ascertained in an outgroup to all of our study populations enables us to compute accurate estimates of the heterozygosity (over a given set of SNPs) throughout
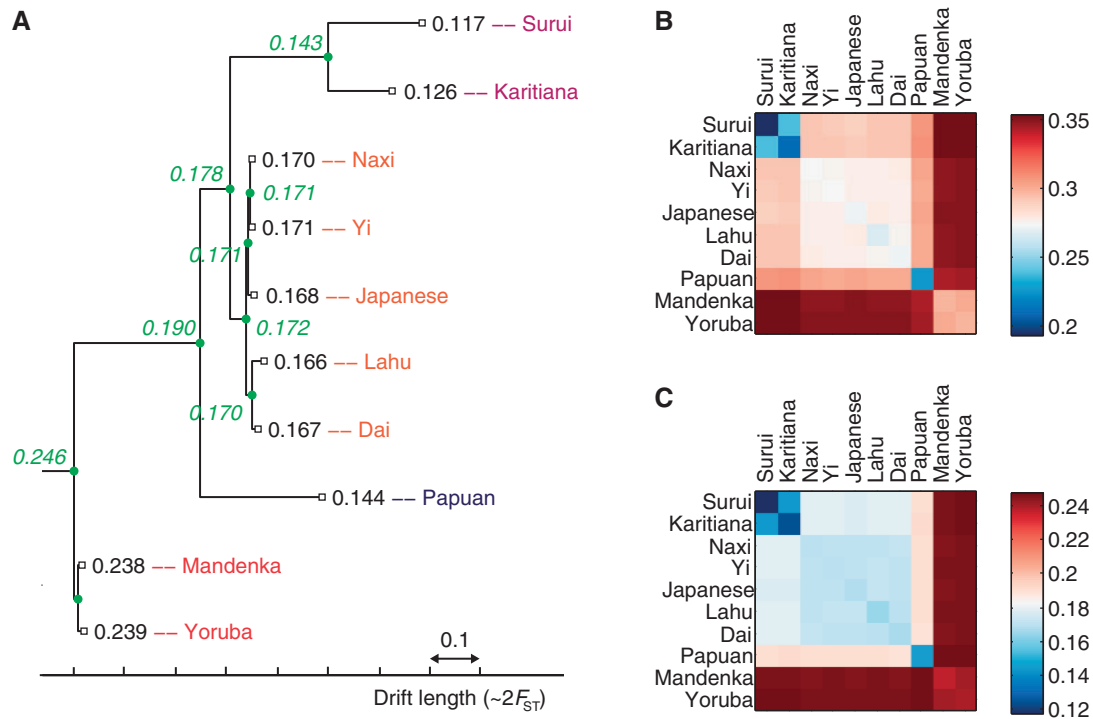
FIG. 6. Ancestral heterozygosity imputed from original Illumina versus San-ascertained SNPs. (A) The 10-population unadmixed tree with estimated average heterozygosities using SNPs from Panel 4 (San ascertainment) of the Affymetrix Human Origins array (Patterson et al. 2012). Numbers in black are direct calculations for modern populations, whereas numbers in green are inferred values at ancestral nodes. (B, C) Computed ancestral heterozygosity at the common ancestor of each pair of modern populations. With unbiased data, values should be equal for pairs having the same common ancestor. (B) Values from a filtered subset of approximately 250,000 SNPs from the published Illumina array data (Li et al. 2008). (C) Values from the Human Origins array excluding SNPs in gene regions.

an unadmixed tree, including at ancestral nodes (see Materials and Methods). This in turn allows us to convert branch lengths from f_2 units to easily interpretable drift lengths (supplementary text S2, Supplementary Material online). In figure 6C, we show our estimates for the heterozygosity (averaged over all San-ascertained SNPs used) at the most recent common ancestor (MRCA) of each pair of present-day populations in the tree. Consensus values are given at the nodes of figure 6A. The imputed heterozygosity should be the same for each pair of populations with the same MRCA, and indeed, with the new data set, the agreement is excellent (fig. 6C). By contrast, inferences of ancestral heterozygosity are much less accurate using HGDP data from the original Illumina SNP array (Li et al. 2008) because of ascertainment bias (fig. 6B); f_2 statistics are also affected but to a lesser degree (supplementary fig. S3, Supplementary Material online), as previously demonstrated (Patterson et al. 2012). We used these heterozygosity estimates to express branch lengths of all of our trees in drift units (supplementary text S2, Supplementary Material online).

Discussion

Comparison with Previous Approaches

The *MixMapper* framework generalizes and automates several previous admixture inference tools based on allele

frequency moment statistics, incorporating them as special cases. Methods such as the 3-population test for admixture and f_4 ratio estimation (Reich et al. 2009; Patterson et al. 2012) have similar theoretical underpinnings, but *MixMapper* provides more extensive information by analyzing more populations simultaneously and automatically considering different tree topologies and sources of gene flow. For example, negative f_3 values—that is, 3-population tests indicating admixture—can be expressed in terms of relationships among f_2 distances between populations in an admixture tree. In general, 3-population tests can be somewhat difficult to interpret because the surrogate ancestral populations may not in fact be closely related to the true participants in the admixture, for example, in the "outgroup case" (Reich et al. 2009; Patterson et al. 2012). The relations among the f_2 statistics incorporate this situation naturally, however, and solving the full system recovers the true branch points wherever they are. As another example, f_4 ratio estimation infers mixture proportions of a single admixture event from f_4 statistics involving the admixed population and four unadmixed populations situated in a particular topology (Reich et al. 2009; Patterson et al. 2012). Whenever data for five such populations are available, the system of all f_2 equations that *MixMapper* solves to obtain the mixture fraction becomes equivalent to the f_4 ratio computation. More importantly, because *MixMapper* infers all of the topological relationships within an admixture tree

automatically by optimizing the solution of the distance equations over all branches, we do not need to specify in advance where the admixture took place—which is not always obvious a priori. By using more than five populations, *MixMapper* also benefits from more data points to constrain the fit.

MixMapper also offers significant advantages over the *qpgraph* admixture tree fitting software (Patterson et al. 2012). Most notably, *qpgraph* requires the user to specify the entire topology of the tree, including admixtures, in advance. This requires either prior knowledge of sources of gene flow relative to the reference populations or a potentially lengthy search to test alternative branch locations. *MixMapper* is also faster and provides the capabilities to convert branch lengths into drift units and to perform bootstrap replicates to measure uncertainty in parameter estimates. Furthermore, *MixMapper* is designed to have more flexible and intuitive input and output and better diagnostics for incorrectly specified models. Although *qpgraph* does fill a niche of fitting very precise models for small sets of populations, it becomes quite cumbersome for more than about seven or eight, whereas *MixMapper* can be run with significantly larger trees without sacrificing efficiency, ease of use, or accuracy of inferences for populations of interest.

Finally, *MixMapper* differs from *TreeMix* (Pickrell and Pritchard 2012) in its emphasis on precise and flexible modeling of individual admixed populations. Stylistically, we view *MixMapper* as "semi-automated" as compared to *TreeMix*, which is almost fully automated. Both approaches have benefits: ours allows more manual guidance and lends itself to interactive use, whereas *TreeMix* requires less user intervention, although some care must be taken in choosing the number of gene flow events to include (10 in the HGDP results shown in supplementary fig. S4, Supplementary Material online) to avoid creating spurious mixtures. With *MixMapper*, we create admixture trees including preselected approximately unadmixed populations together with admixed populations of interest, which are added on a case-by-case basis only if they fit reliably as two- or three-way admixtures. In contrast, *TreeMix* returns a single large-scale admixture tree containing all populations in the input data set, which may include some that can be shown to be admixed by other means but are not modeled as such. Thus, these populations might not be placed well on the tree, which in turn could affect the accuracy of the inferred admixture events. Likewise, the populations ultimately modeled as admixed are initially included as part of an unadmixed tree, where (presumably) they do not fit well, which could introduce errors in the starting tree topology that impact the final results.

Indeed, these methodological differences can be seen to affect inferences for both simulated and real data. For our second simulated admixture tree, *MixMapper* very accurately fit the populations with complicated histories (meant to mimic European and Middle Eastern populations), whereas *TreeMix* only recovered portions of the true tree and also added two inaccurate mixtures (fig. 3). We believe *TreeMix* was hindered in this case by attempting to fit all of the

populations simultaneously and by starting with all of them in an unadmixed tree. In particular, once pop9 (with the lowest proportion of pop7-related admixture) was placed on the unadmixed tree, it likely became difficult to detect as admixed, while pop8's initial placement higher up the tree was likely due to its pop2-related admixture but then obscured this signal in the mixture-fitting phase. Finally, the initial tree shape made populations 3–10 appear to be unequally drifted. Meanwhile, with the HGDP data (figs. 4 and supplementary fig. S4, Supplementary Material online), both methods fit Palestinian, Bedouin, Druze, Mozabite, Uygur, and Hazara as admixed, but *MixMapper* analysis suggested that these populations are better modeled as three-way admixed. *TreeMix* alone fit Brahui, Makrani, Cambodian, and Maya—all of which the 3-population test identifies as admixed but we were unable to place reliably with *MixMapper*—whereas *MixMapper* alone confidently fit Daur, Hezhen, Oroqen, Yakut, Melanesian, and Han. Perhaps most notably, *MixMapper* alone inferred widespread ancient admixture for Europeans; the closest possible signal of such an event in the *TreeMix* model is a migration edge from an ancestor of Native Americans to Russians. We believe that, as in the simulations, *MixMapper* is better suited to finding a common, ancient admixture signal in a group of populations, and more generally to disentangling complex admixture signals from within a large set of populations, and hence it is able to detect admixture in Europeans when *TreeMix* does not.

To summarize, *MixMapper* offers a suite of features that make it better suited than existing methods for the purpose of inferring accurate admixture parameters in data sets containing many specific populations of interest. Our approach provides a middle ground between *qpgraph*, which is designed to fit small numbers of populations within almost no residual errors, and *TreeMix*, which generates large trees with little manual intervention but may be less precise in complex admixture scenarios. Moreover, *MixMapper*'s speed and interactive design allow the user to evaluate the uncertainty and robustness of results in ways that we have found to be very useful (e.g., by comparing two- vs. three-way admixture models or results obtained using alternative scaffold trees).

Ancient European Admixture

Due in part to the flexibility of the *MixMapper* approach, we were able to obtain the notable result that all European populations in the HGDP are best modeled as mixtures between a population related to the common ancestor of Native Americans and a population related to the common ancestor of all non-African populations in our scaffold tree, confirming and extending an admixture signal first reported by Patterson et al. (2012). Our interpretation is that most if not all modern Europeans are descended from at least one large-scale ancient admixture event involving, in some combination, at least one population of Mesolithic European hunter–gatherers; Neolithic farmers, originally from the Near East; and/or other migrants from northern or Central Asia. Either the first or second of these could be related to the "ancient

western Eurasian" branch in figure 5, and either the first or third could be related to the "ancient northern Eurasian" branch. Present-day Europeans differ in the amount of drift they have experienced since the admixture and in the proportions of the ancestry components they have inherited, but their overall profiles are similar.

Our results for Europeans are consistent with several previously published lines of evidence (Pinhasi et al. 2012). First, it has long been hypothesized, based on analysis of a few genetic loci (especially on the Y chromosome), that Europeans are descended from ancient admixtures (Semino et al. 2000; Dupanloup et al. 2004; Soares et al. 2010). Our results also suggest an interpretation for a previously unexplained *frappe* analysis of worldwide human population structure (using $K = 4$ clusters) showing that almost all Europeans contain a small fraction of American-related ancestry (Li et al. 2008). Finally, sequencing of ancient DNA has revealed substantial differentiation in Neolithic Europe between farmers and hunter–gatherers (Bramanti et al. 2009), with the former more closely related to present-day Middle Easterners (Haak et al. 2010) and southern Europeans (Keller et al. 2012; Skoglund et al. 2012) and the latter more similar to northern Europeans (Skoglund et al. 2012), a pattern perhaps reflected in our observed northwest-southeast cline in the proportion of "ancient northern Eurasian" ancestry (table 1). Further analysis of ancient DNA may help shed more light on the sources of ancestry of modern Europeans (Der Sarkissian et al. 2013).

One important new insight of our European analysis is that we detect the same signal of admixture in Sardinian and Basque as in the rest of Europe. As discussed earlier, unlike other Europeans, Sardinian and Basque cannot be confirmed to be admixed using the 3-population test (as in Patterson et al. [2012]), likely due to a combination of less "ancient northern Eurasian" ancestry and more genetic drift since the admixture (table 1). The first point is further complicated by the fact that we have no unadmixed "ancient western Eurasian" population available to use as a reference; indeed, Sardinians themselves are often taken to be such a reference. However, *MixMapper* uncovered strong evidence for admixture in Sardinian and Basque through additivity-checking in the first phase of the program and automatic topology optimization in the second phase, discovering the correct arrangement of unadmixed populations and enabling admixture parameter inference, which we then verified directly with f_4 ratio estimation. Perhaps, the most convincing evidence of the robustness of this finding is that *MixMapper* infers branch points for the ancestral mixing populations that are very similar to those of other Europeans (table 1), a concordance that is most parsimoniously explained by a shared history of ancient admixture among Sardinian, Basque, and other European populations. Finally, we note that because we fit all European populations without assuming Sardinian or Basque to be an unadmixed reference, our estimates of the "ancient northern Eurasian" ancestry proportions in Europeans are larger than those in Patterson et al. (2012) and we believe more accurate than others previously reported (Skoglund et al. 2012).

Future Directions

It is worth noting that of the 52 populations (excluding San) in the HGDP data set, there were 22 that we were unable to fit in a reasonable way either on the unadmixed tree or as admixtures. In part, this was because our instantaneous-admixture model is intrinsically limited in its ability to capture complicated population histories. Most areas of the world have surely witnessed ongoing low levels of interpopulation migration over time, especially between nearby populations, making it difficult to fit admixture trees to the data. We also found cases where having data from more populations would help the fitting process, for example, for three-way admixed populations such as Maya where we do not have a sampled group with a simpler admixture history that could be used to represent two of the three components. Similarly, we found that while Central Asian populations such as Burusho, Pathan, and Sindhi have clear signals of admixture from the three-population test, their ancestry can likely be traced to several different sources (including sub-Saharan Africa in some instances), making them difficult to fit with *MixMapper*, particularly using the HGDP data. Finally, we have chosen here to disregard admixture with archaic humans, which is known to be a small but noticeable component for most populations in the HGDP (Green et al. 2010; Reich et al. 2010). In the future, it will be interesting to extend *MixMapper* and other admixture tree-fitting methods to incorporate the possibilities of multiple-wave and continuous admixture.

In certain applications, full genome sequences are beginning to replace more limited genotype data sets such as ours, but we believe that our methods and SNP-based inference in general will still be valuable in the future. Despite the improving cost-effectiveness of sequencing, it is still much easier and less expensive to genotype samples using a SNP array, and with over 100,000 loci, the data used in this study provide substantial statistical power. Additionally, sequencing technology is currently more error prone, which can lead to biases in allele frequency-based statistics (Pool et al. 2010). We expect that *MixMapper* will continue to contribute to an important toolkit of population history inference methods based on SNP allele frequency data.

Materials and Methods

Model Assumptions and *f*-Statistics

We assume that all SNPs are neutral, biallelic, and autosomal, and that divergence times are short enough that there are no double mutations at a locus. Thus, allele frequency variation—the signal that we harness—is governed entirely by genetic drift and admixture. We model admixture as a one-time exchange of genetic material: two parent populations mix to form a single descendant population whose allele frequencies are a weighted average of the parents'. This model is of course an oversimplification of true mixture events, but it is flexible enough to serve as a first-order approximation.

Our point-admixture model is amenable to allele frequency moment analyses based on *f*-statistics (Reich et al.

2009; Patterson et al. 2012). We primarily make use of the statistic $f_2(A, B) := E_S[(p_A - p_B)^2]$, where p_A and p_B are allele frequencies in populations A and B, and E_S denotes the mean over all SNPs. Expected values of f_2 can be written in terms of admixture tree parameters as described in supplementary text S1, Supplementary Material online. Linear combinations of f_2 statistics can also be used to form the quantities $f_3(C; A, B) := E_S[(p_C - p_A)(p_C - p_B)]$ and $f_4(A, B; C, D) := E_S[(p_A - p_B)(p_C - p_D)]$, which form the bases of the 3- and 4-population tests for admixture, respectively. For all of our f-statistic computations, we use previously described unbiased estimators (Reich et al. 2009; Patterson et al. 2012).

Constructing an Unadmixed Scaffold Tree

Our *MixMapper* admixture-tree-building procedure consists of two phases (fig. 1), the first of which selects a set of unadmixed populations to use as a scaffold tree. We begin by computing f_3 statistics (Reich et al. 2009; Patterson et al. 2012) for all triples of populations P_1, P_2, P_3 in the data set and removing those populations P_3 with any negative values $f_3(P_3; P_1, P_2)$, which indicate admixture. We then use pairwise f_2 statistics to build neighbor-joining trees on subsets of the remaining populations. In the absence of admixture, f_2 distances are additive along paths on a phylogenetic tree (supplementary text S1, Supplementary Material online; cf. Patterson et al. [2012]), meaning that neighbor-joining should recover a tree with leaf-to-leaf distances that are completely consistent with the pairwise f_2 data (Saitou and Nei 1987). However, with real data, the putative unadmixed subsets are rarely completely additive, meaning that the fitted neighbor-joining trees have residual errors between the inferred leaf-to-leaf distances and the true f_2 statistics. These deviations from additivity are equivalent to non-zero results from the four-population test for admixture (Reich et al. 2009; Patterson et al. 2012). We therefore evaluate the quality of each putative unadmixed tree according to its maximum error between fitted and actual pairwise distances: for a tree T having distances d between populations P and Q, the deviation from additivity is defined as $\max\{|d(P, Q) - f_2(P, Q)| : P, Q \in T\}$. *MixMapper* computes this deviation on putatively unadmixed subsets of increasing size, retaining a user-specified number of best subsets of each size in a "beam search" procedure to avoid exponential complexity.

Because of model violations in real data, trees built on smaller subsets are more additive, but they are also less informative; in particular, it is beneficial to include populations from as many continental groups as possible in order to provide more potential branch points for admixture fitting. *MixMapper* provides a ranking of the most additive trees of each size as a guide from which the user chooses a suitable unadmixed scaffold. Once the rank-list of trees has been generated, subject to some constraints (e.g., certain populations required), the user can scan the first several most additive trees for a range of sizes, looking for a balance between coverage and accuracy. This can also be accomplished

by checking whether removing a population from a proposed tree results in a substantial additivity benefit; if so, it may be wise to eliminate it. Similarly, if the population removed from the tree can be modeled well as admixed using the remaining portion of the scaffold, this provides evidence that it should not be part of the unadmixed tree. Finally, *MixMapper* adjusts the scaffold tree that the user ultimately selects by re-optimizing its branch lengths (maintaining the topology inferred from neighbor-joining) to minimize the sum of squared errors of all pairwise f_2 distances.

Within the above guidelines, users should choose the scaffold tree most appropriate for their purposes, which may involve other considerations. In addition to additivity and overall size, it is sometimes desirable to select more or fewer populations from certain geographical, linguistic, or other categories. For example, including a population in the scaffold that is actually admixed might not affect the inferences as long as it is not too closely related to the admixed populations being modeled. At the same time, it can be useful to have more populations in the scaffold around the split points for an admixed population of interest to obtain finer resolution on the branch positions of the mixing populations. For human data in particular, the unadmixed scaffold is only a modeling device; the populations it contains likely have experienced at least a small amount of mixture. A central goal in building the scaffold is to choose populations such that applying this model will not interfere with the conclusions obtained using the program. The interactive design of *MixMapper* allows the user to tweak the scaffold tree very easily to check robustness, and in our analyses, conclusions are qualitatively unchanged for different scaffolds (supplementary figs. S1 and S2 and tables S2–S4, Supplementary Material online).

Two-Way Admixture Fitting

The second phase of *MixMapper* begins by attempting to fit additional populations independently as simple two-way admixtures between branches of the unadmixed tree (fig. 1). For a given admixed population, assuming for the moment that we know the branches from which the ancestral mixing populations split, we can construct a system of equations of f_2 statistics that allows us to infer parameters of the mixture (supplementary text S1, Supplementary Material online). Specifically, the squared allele frequency divergence $f_2(M, X')$ between the admixed population M and each unadmixed population X' can be expressed as an algebraic combination of known branch lengths along with four unknown mixture parameters: the locations of the split points on the two parental branches, the combined terminal branch length, and the mixture fraction (fig. 2A). To solve for the four unknowns, we need at least four unadmixed populations X' that produce a system of four independent constraints on the parameters. This condition is satisfied if and only if the data set contains two populations X'_1 and X'_2 that branch from different points along the lineage connecting the divergence points of the parent populations from the unadmixed tree (supplementary text S1, Supplementary Material online). If the unadmixed tree

contains $n > 4$ populations, we obtain a system of n equations in the four unknowns that in theory is dependent. In practice, the equations are in fact slightly inconsistent because of noise in the f_2 statistics and error in the point-admixture model, so we perform least-squares optimization to solve for the unknowns; having more populations helps reduce the impact of noise.

Algorithmically, *MixMapper* performs two-way admixture fitting by iteratively testing each pair of branches of the unadmixed tree as possible sources of the two ancestral mixing populations. For each choice of branches, *MixMapper* builds the implied system of equations and finds the least-squares solution (under the constraints that unknown branch lengths are nonnegative and the mixture fraction α is between 0 and 1), ultimately choosing the pair of branches and mixture parameters producing the smallest residual norm. Our procedure for optimizing each system of equations uses the observation that upon fixing α, the system becomes linear in the remaining three variables (supplementary text S1, Supplementary Material online). Thus, we can optimize the system by performing constrained linear least squares within a basic one-parameter optimization routine over $\alpha \in [0, 1]$. To implement this approach, we applied MATLAB's `lsqlin` and `fminbnd` functions with a few auxiliary tricks to improve computational efficiency (detailed in the code).

Three-Way Admixture Fitting

MixMapper also fits three-way admixtures, that is, those for which one parent population is itself admixed (fig. 2B). Explicitly, after an admixed population M_1 has been added to the tree, *MixMapper* can fit an additional user-specified admixed population M_2 as a mixture between the M_1 terminal branch and another (unknown) branch of the unadmixed tree. The fitting algorithm proceeds in a manner analogous to the two-way mixture case: *MixMapper* iterates through each possible choice of the third branch, optimizing each implied system of equations expressing f_2 distances in terms of mixture parameters. With two admixed populations, there are now $2n + 1$ equations, relating observed values of $f_2(M_1, X')$ and $f_2(M_2, X')$ for all unadmixed populations X', and also $f_2(M_1, M_2)$, to eight unknowns: two mixture fractions, α_1 and α_2, and six branch length parameters (fig. 2B). Fixing α_1 and α_2 results in a linear system as before, so we perform the optimization using MATLAB's `lsqlin` within `fminsearch` applied to α_1 and α_2 in tandem. The same mathematical framework could be extended to optimizing the placement of populations with arbitrarily many ancestral admixture events, but for simplicity and to reduce the risk of overfitting, we chose to limit this version of *MixMapper* to three-way admixtures.

Expressing Branch Lengths in Drift Units

All of the tree-fitting computations described thus far are performed using pairwise distances in f_2 units, which are mathematically convenient to work with owing to their additivity along a lineage (in the absence of admixture). However, f_2 distances are not directly interpretable in the

same way as genetic drift D, which is a simple function of time and population size:

$$D \approx 1 - \exp(-t/2N_e) \approx 2 \cdot F_{ST},$$

where t is the number of generations and N_e is the effective population size (Nei 1987). To convert f_2 distances to drift units, we apply a new formula, dividing twice the f_2-length of each branch by the heterozygosity value that we infer for the ancestral population at the top of the branch (supplementary text S2, Supplementary Material online). Qualitatively speaking, this conversion corrects for the relative stretching of f_2 branches at different portions of the tree as a function of heterozygosity (Patterson et al. 2012). To infer ancestral heterozygosity values accurately, it is critical to use SNPs that are ascertained in an outgroup to the populations involved, which we address later.

Before inferring heterozygosities at ancestral nodes of the unadmixed tree, we must first determine the location of the root (which is neither specified by neighbor-joining nor involved in the preceding analyses). *MixMapper* does so by iterating through branches of the unadmixed tree, temporarily rooting the tree along each branch, and then checking for consistency of the resulting heterozygosity estimates. Explicitly, for each internal node P, we split its present-day descendants (according to the re-rooted tree) into two groups G_1 and G_2 according to which child branch of P they descend from. For each pair of descendants, one from G_1 and one from G_2, we compute an inferred heterozygosity at P (supplementary text S2, Supplementary Material online). If the tree is rooted properly, these inferred heterozygosities are consistent, but if not, there exist nodes P for which the heterozygosity estimates conflict. *MixMapper* thus infers the location of the root as well as the ancestral heterozygosity at each internal node, after which it applies the drift length conversion as a postprocessing step on fitted f_2 branch lengths.

Bootstrapping

To measure the statistical significance of our parameter estimates, we compute bootstrap confidence intervals (Efron 1979; Efron and Tibshirani 1986) for the inferred branch lengths and mixture fractions. Our bootstrap procedure is designed to account for both the randomness of the drift process at each SNP and the random choice of individuals sampled to represent each population. First, we divide the genome into 50 evenly sized blocks, with the premise that this scale should easily be larger than that of linkage disequilibrium among our SNPs. Then, for each of 500 replicates, we resample the data set by 1) selecting 50 of these SNP blocks at random with replacement; and 2) for each population group, selecting a random set of individuals with replacement, preserving the number of individuals in the group.

For each replicate, we recalculate all pairwise f_2 distances and present-day heterozygosity values using the resampled SNPs and individuals (adjusting the bias-correction terms to account for the repetition of individuals) and then construct the admixture tree of interest. Even though the mixture parameters we estimate—branch lengths and mixture

fractions—depend in complicated ways on many different random variables, we can directly apply the nonparametric bootstrap to obtain confidence intervals (Efron and Tibshirani 1986). For simplicity, we use a percentile bootstrap; thus, our 95% confidence intervals indicate 2.5 and 97.5 percentiles of the distribution of each parameter among the replicates.

Computationally, we parallelize *MixMapper*'s mixture-fitting over the bootstrap replicates using MATLAB's Parallel Computing Toolbox.

Evaluating Fit Quality

When interpreting admixture inferences produced by methods such as *MixMapper*, it is important to ensure that best-fit models are in fact accurate. Although formal tests for goodness of fit do not generally exist for methods of this class, we use several criteria to evaluate the mixture fits produced by *MixMapper* and distinguish high-confidence results from possible artifacts of overfitting or model violations.

First, we can compare *MixMapper* results to information obtained from other methods, such as the 3-population test (Reich et al. 2009; Patterson et al. 2012). Negative f_3 values indicate robustly that the tested population is admixed, and comparing f_3 statistics for different reference pairs can give useful clues about the ancestral mixing populations. Thus, while the three-population test relies on similar data to *MixMapper*, its simpler form makes it useful for confirming that *MixMapper* results are reasonable.

Second, the consistency of parameter values over bootstrap replicates gives an indication of the robustness of the admixture fit in question. All results with real data have some amount of associated uncertainty, which is a function of sample sizes, SNP density, intrapopulation homogeneity, and other aspects of the data. Given these factors, we place less faith in results with unexpectedly large error bars. Most often, this phenomenon is manifested in the placement of ancestral mixing populations: for poorly fitting admixtures, branch choices often change from one replicate to the next, signaling unreliable results.

Third, we find that results where one ancestral population is very closely related to the admixed population and contributes more than 90% of the ancestry are often unreliable. We expect that if we try to fit a nonadmixed population as an admixture, *MixMapper* should return a closely related population as the first branch with mixture fraction $\alpha \approx 1$ (and an arbitrary second branch). Indeed, we often observe this pattern in the context of verifying that certain populations make sense to include in the scaffold tree. Further evidence of overfitting comes when the second ancestry component, which contributes only a few percent, either bounces from branch to branch over the replicates, is located at the very tip of a leaf branch, or is historically implausible.

Fourth, for any inferred admixture event, the two mixing populations must be contemporaneous. As we cannot resolve the three pieces of terminal drift lengths leading to admixed populations (fig. 2A) and our branch lengths depend both on population size and absolute time, we

cannot say for sure whether this property is satisfied for any given mixture fit. In some cases, however, it is clear that no realization of the variables could possibly be consistent: for example, if we infer an admixture between a very recent branch and a very old one with a small value of the total mixed drift—and hence the terminal drift c—then we can confidently say the mixture is unreasonable.

Finally, when available, we also use prior historical or other external knowledge to guide what we consider to be reasonable. Sometimes, the model that appears to fit the data best has implications that are clearly historically implausible; often when this is true one or more of the evaluation criteria listed earlier can be invoked as well. Of course, the most interesting findings are often those that are new and surprising, but we subject such results to an extra degree of scrutiny.

Data Set and Ascertainment

We analyzed a SNP data set from 934 HGDP individuals grouped in 53 populations (Rosenberg et al. 2002; Li et al. 2008). Unlike most previous studies of the HGDP samples, however, we worked with recently published data generated using the new Affymetrix Axiom Human Origins Array (Patterson et al. 2012), which was designed with a simple ascertainment scheme for accurate population genetic inference (Keinan et al. 2007). It is well known that ascertainment bias can cause errors in estimated divergences among populations (Clark et al. 2005; Albrechtsen et al. 2010), as choosing SNPs based on their properties in modern populations induces nonneutral spectra in related samples. Although there do exist methods to correct for ascertainment bias (Nielsen et al. 2004), it is much more desirable to work with a priori bias-free data, especially given that typical SNP arrays are designed using opaque ascertainment schemes.

To avoid these pitfalls, we used panel 4 of the new array, which consists of 163,313 SNPs that were ascertained as heterozygous in the genome of a San individual (Keinan et al. 2007). This panel is special because there is evidence that the San are approximately an outgroup to all other modern-day human populations (Li et al. 2008; Gronau et al. 2011). Thus, while the panel 4 ascertainment scheme distorts the San allele frequency spectrum, it is nearly neutral with respect to all other populations. In other words, we can think of the ascertainment as effectively choosing a set of SNPs (biased toward San heterozygosity) at the common ancestor of the remaining 52 populations, after which drift occurs in a bias-free manner. We excluded 61,369 SNPs that are annotated as falling between the transcription start site and end site of a gene in the UCSC Genome Browser database (Fujita et al. 2011). Most of the excluded SNPs are not within actual exons, but as expected, the frequency spectra at these "gene region" loci were slightly shifted toward fixed classes relative to other SNPs, indicative of the action of selection (supplementary fig. S5, Supplementary Material online). As we assume neutrality in all of our analyses, we chose to remove these SNPs.

Simulations

Our first simulated tree was generated using the `ms` (Hudson 2002) command

```
ms 350 500 -t 50 -r 99.9998 500000 -I 7 50 50
50 50 50 50 50 -n 7 2 -n 1 2 -n 2 2 -ej 0.04 2 1 -es
0.02 6 0.4 -ej 0.06 6 3 -ej 0.04 8 5 -ej 0.08 5 4
-ej 0.12 4 3 -ej 0.2 3 1 -ej 0.3 1 7 -en 0.37 1.
```

After ascertainment, we used a total of 95,997 SNPs.

Our second simulated tree was generated with the command

```
ms 550 500 -t 50 -r 99.9998 500000 -I 11 50 50
50 50 50 50 50 50 50 50 50 -n 11 2 -n 1 2 -n 2 2
-em 0.002 4 3 253.8 -em 0.004 4 3 0 -es 0.002
8 0.2 -en 0.002 8 2 -ej 0.02 8 2 -ej 0.02 4 5
-ej 0.04 2 1 -ej 0.04 5 3 -es 0.04 12 0.4 -es
0.04 9 0.2 -em 0.042 10 9 253.8 -em 0.044 10 9
0 -ej 0.06 12 7 -ej 0.06 9 7 -ej 0.06 14 10 -ej
0.06 13 10 -ej 0.08 7 6 -ej 0.12 6 3 -ej 0.16
10 3 -ej 0.2 3 1 -ej 0.3 1 11 -en 0.3 11 1.
```

After ascertainment, we used a total of 96,258 SNPs. When analyzing this data set in *TreeMix*, we chose to fit a total of five admixtures based on the residuals of the pairwise distances (maximum of approximately three standard errors) and our knowledge that this is the number in the true admixture tree (to make for a fair comparison).

Software

Source code for the *MixMapper* software is available at http://groups.csail.mit.edu/cb/mixmapper/ (last accessed June 14, 2013).

Supplementary Material

Supplementary figures S1–S5, tables S1–S5, and text S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

References

Albrechtsen A, Nielsen F, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol.* 27: 2534–2547.

Bramanti B, Thomas M, Haak W, et al. (11 co-authors). 2009. Genetic discontinuity between local hunter-gatherers and Central Europe's first farmers. *Science* 326:137–140.

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

Chikhi L, Bruford M, Beaumont M. 2001. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* 158:1347–1362.

Clark A, Hubisz M, Bustamante C, Williamson S, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15:1496–1502.

Der Sarkissian C, Balanovsky O, Brandt G, et al. (11 co-authors). 2013. Ancient DNA reveals prehistoric gene-flow from Siberia in the complex human population history of North East Europe. *PLoS Genet.* 9:e1003296.

Dupanloup I, Bertorelle G, Chikhi L, Barbujani G. 2004. Estimating the impact of prehistoric admixture on the genome of Europeans. *Mol Biol Evol.* 21:1361–1372.

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann Stat.* 7:1–26.

Efron B, Tibshirani R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci.* 1:54–75.

Fujita P, Rhead B, Zweig A, et al. (11 co-authors). 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* 39: D876–D882.

Gravel S, Henn B, Gutenkunst R, et al. (11 co-authors). 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A.* 108:11983–11988.

Green R, Krause J, Briggs A, et al. (11 co-authors). 2010. A draft sequence of the Neandertal genome. *Science* 328:710–722.

Gronau I, Hubisz M, Gulko B, Danko C, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet.* 43:1031–1034.

Haak W, Balanovsky O, Sanchez J, et al. (11 co-authors). 2010. Ancient DNA from European early Neolithic farmers reveals their Near Eastern affinities. *PLoS Biol.* 8:e1000536.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

HUGO Pan-Asian SNP Consortium. 2009. Mapping human genetic diversity in Asia. *Science* 326:1541–1545.

Huson D, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.

Keinan A, Mullikin J, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet.* 39:1251–1255.

Keller A, Graefen A, Ball M, et al. (11 co-authors). 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun.* 3:698.

Laval G, Patin E, Barreiro L, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5: e10284.

Li J, Absher D, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193:1233–1254.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nielsen R, Hubisz M, Clark A. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168:2373–2382.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192:1065–1093.

Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8: e1002967.

Pinhasi R, Thomas M, Hofreiter M, Currat M, Burger J. 2012. The genetic history of Europeans. *Trends Genet.* 28:496–505.

Pool J, Hellmann I, Jensen J, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Res.* 20:291–300.

Reich D, Green R, Kircher M, et al. (11 co-authors). 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1053–1060.

Reich D, Patterson N, Kircher M, et al. (11 co-authors). 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet.* 89:516–528.

MBE

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

Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, Feldman M. 2002. Genetic structure of human populations. *Science* 298:2381–2385.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4: 406–425.

Semino O, Passarino G, Oefner P, et al. (11 co-authors). 2000. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science* 290:1155–1159.

Sirén J, Marttinen P, Corander J. 2011. Reconstructing population histories from single nucleotide polymorphism data. *Mol Biol Evol.* 28: 673–683.

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

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt H, Torroni A, Richards M. 2010. The archaeogenetics of Europe. *Curr Biol.* 20: R174–R183.

Sousa V, Fritz M, Beaumont M, Chikhi L. 2009. Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* 181:1507–1519.

Wall J, Lohmueller K, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol.* 26:1823–1827.

Wang J. 2003. Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* 164:747–765.

Xu S, Pugach I, Stoneking M, et al. (6 co-authors). 2012. Genetic dating indicates that the Asian–Papuan admixture through eastern Indonesia corresponds to the Austronesian expansion. *Proc Natl Acad Sci U S A.* 109:4574–4579.

Yu Y, Degnan J, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.* 8:e1002660.