Supplementary Materials: Efficient moment-based inference of admixture parameters and sources of gene flow

Mark Lipson, Po-Ru Loh, Alex Levin, David Reich, Nick Patterson, and Bonnie Berger

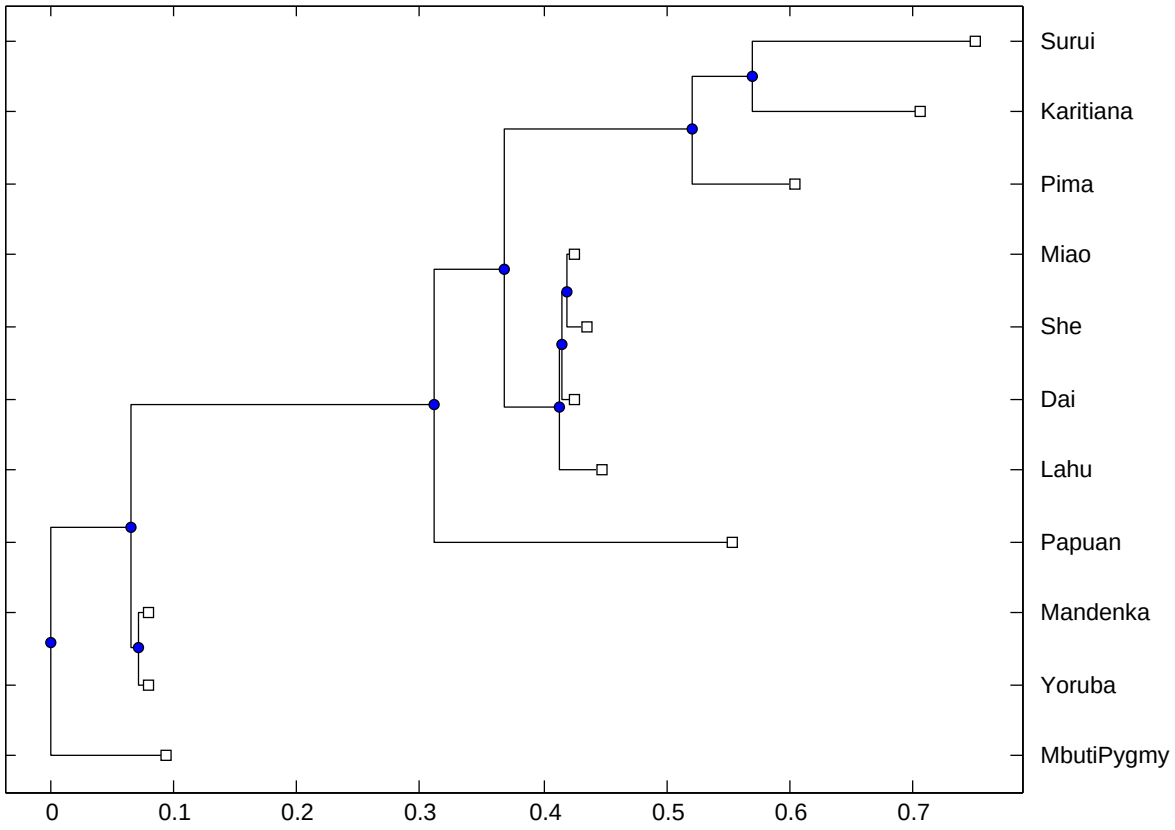

Figure S1. Alternative scaffold tree with 11 populations used to evaluate robustness of results to scaffold choice. We included Mbuti Pygmy, who are known to be admixed, to help demonstrate that *MixMapper* inferences are robust to deviations from additivity in the scaffold; see Tables S2–S4 for full results. Distances are in drift units.

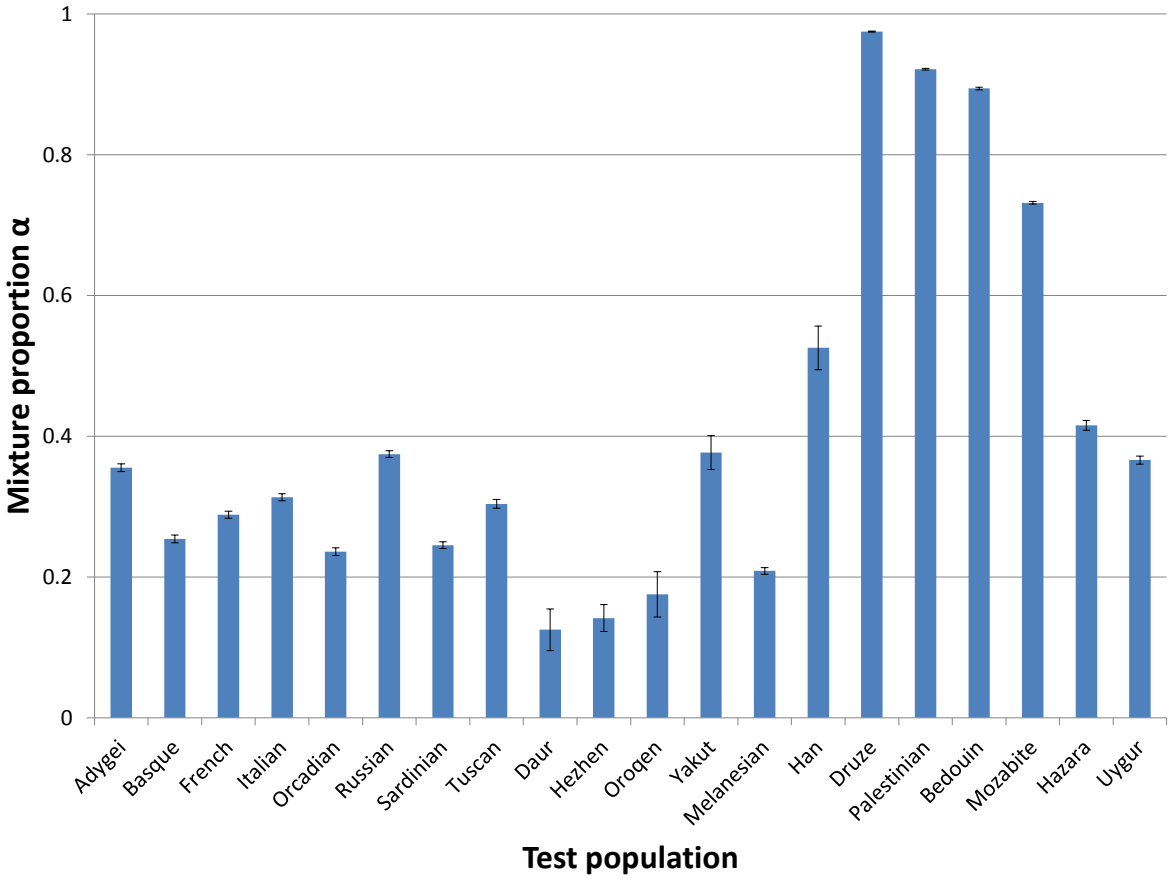

Figure S2. Summary of mixture proportions α inferred with alternative 9-population scaffold trees. We ran *MixMapper* for all 20 admixed test populations using nine different scaffold trees obtained by removing each population except Papuan one at a time from our full 10-population scaffold. (Papuan is needed to maintain continental representation.) For each test population and each scaffold, we recorded the median bootstrap-inferred value of α over all replicates having branching patterns similar to the primary topology. Shown here are the means and standard deviations of the nine medians. In all cases, α refers to the proportion of ancestry from the first branch as in Tables 1–3.

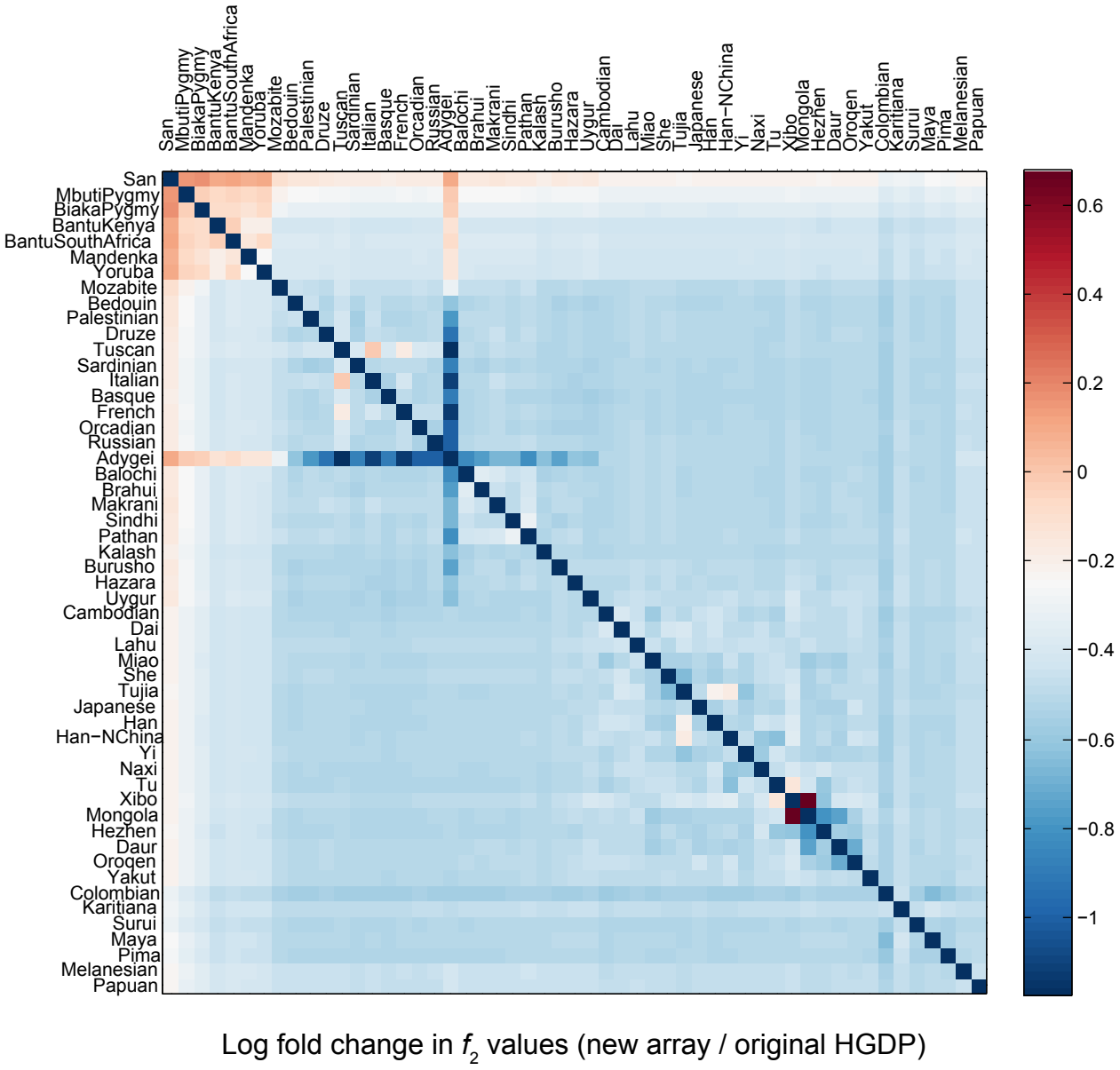

Figure S3. Comparison of f_2 distances computed using original Illumina vs. San-ascertained SNPs. The heat map shows the log fold change in f_2 values obtained from the original HGDP data (Li et al., 2008) versus the San-ascertained data (Patterson et al., 2012) used in this study.

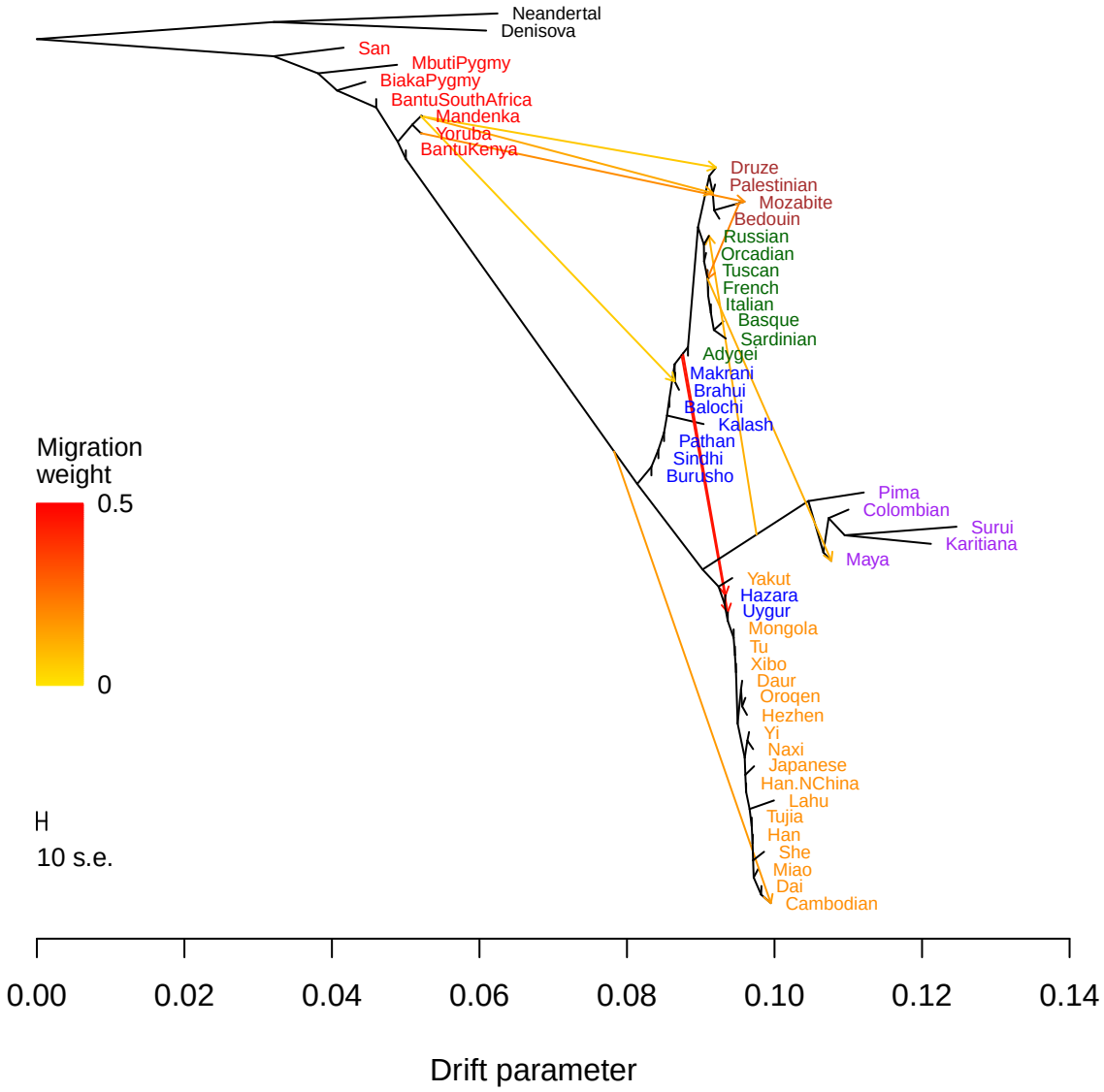

Figure S4. *TreeMix* results on the HGDP. Admixture graph for HGDP populations obtained with the *TreeMix* software, as reported in Pickrell and Pritchard (2012). Figure is reproduced from Pickrell and Pritchard (2012) with permission of the authors and under the Creative Commons Attribution License.

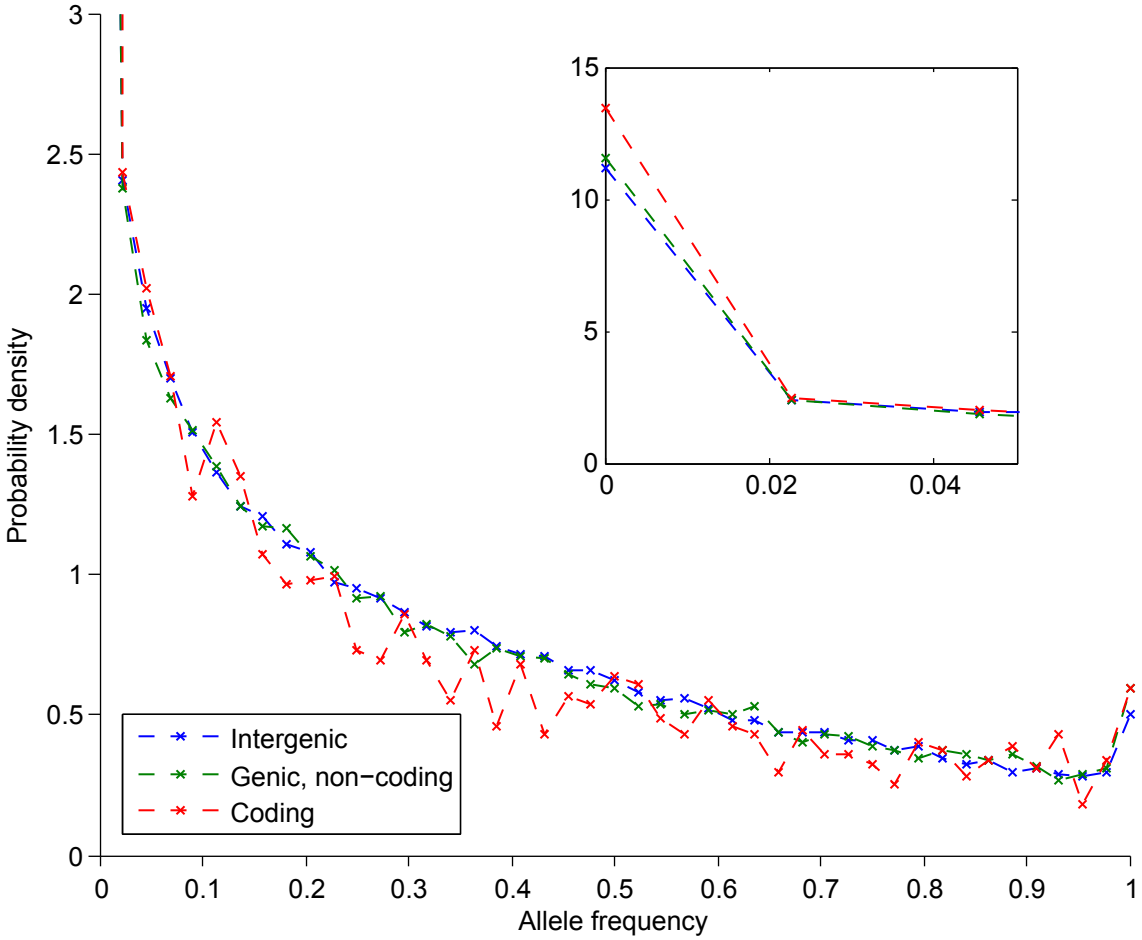

Figure S5. Comparison of allele frequency spectra within and outside gene regions. We divided the Panel 4 (San-ascertained) SNPs into three groups: those outside gene regions (101,944), those within gene regions but not in exons (58,110), and those within coding regions (3259). Allele frequency spectra restricted to each group are shown for the Yoruba population. Reduced heterozygosity within exon regions is evident, which suggests the action of purifying selection. (Inset) We observe the same effect in the genic, non-coding spectrum; it is less noticeable but can be seen at the edge of the spectrum.

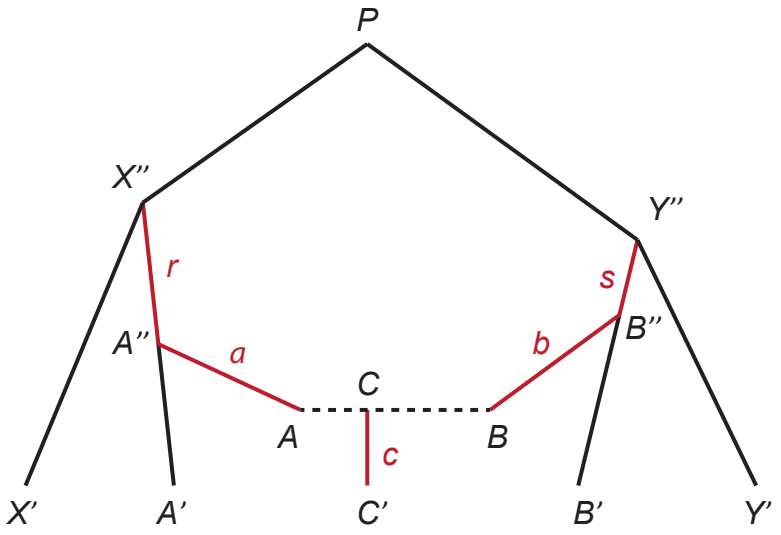

Figure S6. Schematic of part of an admixture tree. Population C is derived from an admixture of populations A and B with proportion α coming from A. The f_2 distances from C' to the present-day populations A', B', X', Y' give four relations from which we are able to infer four parameters: the mixture fraction α, the locations of the split points A'' and B'' (i.e., r and s), and the combined drift $\alpha^2 a + (1-\alpha)^2 b + c$.

Table S1. Mixture parameters for simulated data.

AdmixedPop	Branch1 + Branch2	# rep	α	Branch1Loc	Branch2Loc	MixedDrift
First tree						
pop6	pop3 + pop5	500	0.253-0.480	0.078-0.195 / 0.214	0.050-0.086 / 0.143	0.056-0.068
pop6 (true)	pop3 + pop5		0.4	0.107 / 0.213	0.077 / 0.145	0.066
Second tree						
pop4	pop3 + pop5	500	0.382-0.652	0.039-0.071 / 0.076	0.032-0.073 / 0.077	0.010-0.020
pop4 (true)	pop3 + pop5		0.4	0.071 / 0.077	0.038 / 0.077	0.016
pop9	Anc3–7 + pop7	490	0.653-0.915	0.048-0.091 / 0.140	0.013-0.134 / 0.147	0.194-0.216
pop9 (true)	Anc3–7 + pop7		0.8	0.077 / 0.145	0.037 / 0.145	0.194
pop10	Anc3–7 + pop7	500	0.502-0.690	0.047-0.091 / 0.140	0.021-0.067 / 0.147	0.151-0.167
pop10 (true)	Anc3–7 + pop7		0.6	0.077 / 0.145	0.037 / 0.145	0.150
AdmixedPop2	AdmixedPop1 + Branch3	# rep	α_2	Branch3Loc		
pop8	pop10 + pop2	304	0.782-0.822	0.007-0.040 / 0.040		
	pop10 + Anc1–2	193	0.578-0.756	0.009-0.104 / 0.148		
pop8 (true)	pop10 + pop2		0.8	0.020 / 0.039		

NOTE.—Mixture parameters inferred by *MixMapper* for simulated data, followed by true values for each simulated admixed population. Branch1 and Branch2 are the optimal split points for the mixing populations, with α the proportion of ancestry from Branch1; topologies are shown that that occur for at least 20 of 500 bootstrap replicates. The mixed drift parameters for the three-way admixed pop8 are not well-defined in the simulated tree and are omitted. The branch "Anc3–7" is the common ancestral branch of pops 3–7, and the branch "Anc1–2" is the common ancestral branch of pops 1–2. See Figure 2 and the caption of Table 1 for descriptions of the parameters and Figure 3 for plots of the results.

Table S2. Mixture parameters for Europeans inferred with an alternative scaffold tree.

AdmixedPop	# rep	α	Branch1Loc (Anc. N. Eurasian)	Branch2Loc (Anc. W. Eurasian)	MixedDrift
Adygei	488	0.278-0.475	0.035-0.078 / 0.151	0.158-0.191 / 0.246	0.078-0.093
Basque	273	0.221-0.399	0.055-0.111 / 0.153	0.164-0.194 / 0.244	0.108-0.124
French	380	0.240-0.410	0.054-0.108 / 0.152	0.165-0.192 / 0.245	0.093-0.106
Italian	427	0.245-0.426	0.047-0.103 / 0.152	0.155-0.188 / 0.246	0.095-0.110
Orcadian	226	0.214-0.387	0.061-0.131 / 0.153	0.174-0.197 / 0.244	0.098-0.116
Russian	472	0.296-0.490	0.047-0.093 / 0.151	0.165-0.197 / 0.246	0.080-0.095
Sardinian	390	0.189-0.373	0.045-0.104 / 0.152	0.160-0.190 / 0.245	0.110-0.125
Tuscan	413	0.238-0.451	0.039-0.096 / 0.152	0.153-0.191 / 0.245	0.093-0.111

NOTE.—Mixture parameters inferred by *MixMapper* for modern-day European populations using an alternative unadmixed scaffold tree containing 11 populations: Yoruba, Mandenka, Mbuti Pygmy, Papuan, Dai, Lahu, Miao, She, Karitiana, Suruí, and Pima (see Figure S1). The parameter estimates are very similar to those obtained with the original scaffold tree (Table 1), with α slightly higher on average. The bootstrap support for the branching position of "ancient northern Eurasian" plus "ancient western Eurasian" is also somewhat lower, with the remaining replicates almost all placing the first ancestral population along the Pima branch instead. However, this is perhaps not surprising given evidence of European-related admixture in Pima; overall, our conclusions are unchanged, and the results appear quite robust to perturbations in the scaffold. See Figure 2A and the caption of Table 1 for descriptions of the parameters.

Table S3. Mixture parameters for other populations modeled as two-way admixtures inferred with an alternative scaffold tree.

AdmixedPop	Branch1 + Branch2	# rep	α	Branch1Loc	Branch2Loc	MixedDrift
Daur	Anc. N. Eurasian + She	264	0.225-0.459	0.005-0.052 / 0.151	0.002-0.014 / 0.016	0.014-0.024
	Anc. N. Eurasian + Miao	213	0.235-0.422	0.005-0.049 / 0.151	0.002-0.008 / 0.008	0.014-0.024
Hezhen	Anc. N. Eurasian + She	257	0.230-0.442	0.005-0.050 / 0.151	0.002-0.010 / 0.016	0.012-0.034
	Anc. N. Eurasian + Miao	217	0.214-0.444	0.005-0.047 / 0.151	0.002-0.008 / 0.008	0.013-0.037
Oroqen	Anc. N. Eurasian + She	336	0.284-0.498	0.010-0.052 / 0.151	0.003-0.015 / 0.016	0.017-0.036
	Anc. N. Eurasian + Miao	149	0.271-0.476	0.007-0.046 / 0.151	0.002-0.008 / 0.008	0.018-0.039
Yakut	Anc. N. Eurasian + Miao	246	0.648-0.864	0.004-0.018 / 0.151	0.005-0.008 / 0.008	0.032-0.043
	Anc. East Asian + Pima	71	0.917-0.973	0.008-0.020 / 0.045	0.022-0.083 / 0.083	0.028-0.042
	Anc. N. Eurasian + She	161	0.664-0.865	0.004-0.018 / 0.151	0.003-0.017 / 0.017	0.030-0.043
Melanesian	Dai + Papuan	331	0.168-0.268	0.009-0.011 / 0.011	0.167-0.204 / 0.246	0.089-0.115
	Lahu + Papuan	78	0.174-0.266	0.005-0.034 / 0.034	0.167-0.203 / 0.244	0.089-0.118
Han	Karitiana + She	167	0.007-0.025	0.026-0.134 / 0.134	0.001-0.006 / 0.016	0.000-0.004
	She + Surui	54	0.971-0.994	0.001-0.006 / 0.016	0.017-0.180 / 0.180	0.000-0.003
	Anc. N. Eurasian + She	65	0.021-0.080	0.004-0.105 / 0.152	0.001-0.007 / 0.016	0.000-0.003
	Pima + She	82	0.009-0.033	0.022-0.085 / 0.085	0.001-0.007 / 0.016	0.000-0.004

NOTE.—Mixture parameters inferred by *MixMapper* for non-European populations fit as two-way admixtures using an alternative unadmixed scaffold tree containing 11 populations: Yoruba, Mandenka, Mbuti Pygmy, Papuan, Dai, Lahu, Miao, She, Karitiana, Suruí, and Pima (see Figure S1). The results for the first four populations are very similar to those obtained with the original scaffold tree, except that α is now estimated to be roughly 20% higher. Melanesian is fit essentially identically as before. Han, however, now appears nearly unadmixed, which we suspect is due to the lack of an appropriate northern East Asian population related to one ancestor (having removed Japanese). See Figure 2A and the caption of Table 1 for descriptions of the parameters; branch choices are shown that that occur for at least 50 of 500 bootstrap replicates. The "Anc. East Asian" branch is the common ancestral branch of the four East Asian populations in the unadmixed tree.

Table S4. Mixture parameters for populations modeled as three-way admixtures inferred with an alternative scaffold tree.

AdmixedPop2	Branch3	# rep	α_2	Branch3Loc	MixedDrift1A	FinalDrift1B	MixedDrift2
Druze	Mandenka	309	0.958-0.984	0.004-0.009 / 0.009	0.088-0.102	0.021-0.029	0.005-0.013
Palestinian	Mandenka	249	0.907-0.935	0.008-0.009 / 0.009	0.087-0.100	0.022-0.030	0.001-0.008
	Anc. W. Eurasian	92	0.822-0.893	0.050-0.122 / 0.246	0.102-0.126	0.000-0.019	0.011-0.023
Bedouin	Mandenka	303	0.852-0.918	0.006-0.009 / 0.009	0.086-0.101	0.022-0.030	0.007-0.019
Mozabite	Mandenka	339	0.684-0.778	0.006-0.009 / 0.009	0.095-0.112	0.010-0.021	0.018-0.032
	Yoruba	50	0.673-0.778	0.005-0.010 / 0.010	0.093-0.111	0.010-0.020	0.018-0.031
Hazara	Anc. East Asian	390	0.350-0.464	0.009-0.023 / 0.045	0.084-0.119	0.001-0.033	0.004-0.012
Uygur	Anc. East Asian	390	0.312-0.432	0.007-0.022 / 0.045	0.091-0.124	0.000-0.027	0.000-0.009

NOTE.—Mixture parameters inferred by *MixMapper* for populations fit as three-way admixtures using an alternative unadmixed scaffold tree containing 11 populations: Yoruba, Mandenka, Mbuti Pygmy, Papuan, Dai, Lahu, Miao, She, Karitiana, Suruí, and Pima (see Figure S1). In all cases one parent population splits from the (admixed) Sardinian branch and the other from Branch3. All the parameters are quite similar to those obtained with the original scaffold with only some relative changes in bootstrap support among alternative topologies. See Figure 2B and the caption of Table 1 for further descriptions of the parameters; branch choices are shown that that occur for at least 50 of the 390 bootstrap replicates having the majority branch choices for the two-way Sardinian fit. The "Anc. East Asian" branch is the common ancestral branch of the four East Asian populations in the unadmixed tree.

Table S5. Mixture proportions for Sardinian and Basque from f_4 ratio estimation.

Test pop.	Asian pop.	American pop.	α
Sardinian	Dai	Karitiana	23.3 $\pm$ 6.3
Sardinian	Dai	Suruí	24.5 $\pm$ 6.7
Sardinian	Lahu	Karitiana	23.1 $\pm$ 7.0
Sardinian	Lahu	Suruí	24.7 $\pm$ 7.6
Basque	Dai	Karitiana	22.8 $\pm$ 7.0
Basque	Dai	Suruí	24.0 $\pm$ 7.6
Basque	Lahu	Karitiana	23.1 $\pm$ 7.4
Basque	Lahu	Suruí	24.7 $\pm$ 8.0

NOTE.—To validate the mixture proportions estimated by *MixMapper* for Sardinian and Basque, we applied f_4 ratio estimation. The fraction α of "ancient northern Eurasian" ancestry was estimated as $\alpha = f_4$(Papuan, Asian; Yoruba, European) / f_4(Papuan, Asian; Yoruba, American), where the European population is Sardinian or Basque, Asian is Dai or Lahu, and American is Karitiana or Suruí. Standard errors are from 500 bootstrap replicates. Note that this calculation assumes the topology of the ancestral mixing populations as inferred by *MixMapper* (Figure 5A).

Text S1. f-statistics and population admixture.

Here we include derivations of the allele frequency divergence equations solved by *MixMapper* to determine the optimal placement of admixed populations. These results were first presented in Reich et al. (2009) and Patterson et al. (2012), and we reproduce them here for completeness, with slightly different emphasis and notation. We also describe in the final paragraph (and in more detail in Material and Methods) how the structure of the equations leads to a particular form of the system for a full admixture tree.

Our basic quantity of interest is the f-statistic f_2, as defined in Reich et al. (2009), which is the squared allele frequency difference between two populations at a biallelic SNP. That is, at SNP locus i, we define

$$f_2^i(A, B) := (p_A - p_B)^2,$$

where p_A is the frequency of one allele in population A and p_B is the frequency of the allele in population B. This is the same as Nei's minimum genetic distance D_{AB} for the case of a biallelic locus (Nei, 1987). As in Reich et al. (2009), we define the unbiased estimator $\hat{f}_2^i(A, B)$, which is a function of finite population samples:

$$\hat{f}_2^i(A, B) := (\hat{p}_A - \hat{p}_B)^2 - \frac{\hat{p}_A(1 - \hat{p}_A)}{n_A - 1} - \frac{\hat{p}_B(1 - \hat{p}_B)}{n_B - 1},$$

where, for each of A and B, $\hat{p}$ is the the empirical allele frequency and n is the total number of sampled alleles.

We can also think of $f_2^i(A, B)$ itself as the outcome of a random process of genetic history. In this context, we define

$$F_2^i(A, B) := E((p_A - p_B)^2),$$

the expectation of $(p_A - p_B)^2$ as a function of population parameters. So, for example, if B is descended from A via one generation of Wright-Fisher genetic drift in a population of size N, then $F_2^i(A, B) = p_A(1 - p_A)/2N$.

While $\hat{f}_2^i(A,B)$ is unbiased, its variance may be large, so in practice, we use the statistic

$$\hat{f}_2(A,B) := \frac{1}{m}\sum_{i=1}^{m} \hat{f}_2^i(A,B),$$

i.e., the average of $\hat{f}_2^i(A,B)$ over a set of m SNPs. As we discuss in more detail in Text S2, $F_2^i(A,B)$ is not the same for different loci, meaning $\hat{f}_2(A,B)$ will depend on the choice of SNPs. However, we do know that $\hat{f}_2(A,B)$ is an unbiased estimator of the true average $f_2(A,B)$ of $f_2^i(A,B)$ over the set of SNPs.

The utility of the f_2 statistic is due largely to the relative ease of deriving equations for its expectation between populations on an admixture tree. The following derivations are borrowed from (Reich et al., 2009). As above, let the frequency of a SNP i in population X be p_X. Then, for example,

$$\begin{aligned} E(f_2^i(A,B)) &= E((p_A - p_B)^2) \\ &= E((p_A - p_P + p_P - p_B)^2) \\ &= E((p_A - p_P)^2) + E((p_P - p_B)^2) + 2E((p_A - p_P)(p_P - p_B)) \\ &= E(f_2^i(A,P)) + E(f_2^i(B,P)), \end{aligned}$$

since the genetic drifts $p_A - p_P$ and $p_P - p_B$ are uncorrelated and have expectation 0. We can decompose these terms further; if Q is a population along the branch between A and P, then:

$$\begin{aligned} E(f_2^i(A,P)) &= E((p_A - p_P)^2) \\ &= E((p_A - p_Q + p_Q - p_P)^2) \\ &= E((p_A - p_Q)^2) + E((p_Q - p_P)^2) + 2E((p_A - p_Q)(p_Q - p_P)) \\ &= E(f_2^i(A,Q)) + E(f_2^i(Q,P)). \end{aligned}$$

Here, again, $E(p_A - p_Q) = E(p_Q - p_P) = 0$, but $p_A - p_Q$ and $p_Q - p_P$ are not independent;

for example, if $p_Q - p_P = -p_P$, i.e. $p_Q = 0$, then necessarily $p_A - p_Q = 0$. However, $p_A - p_Q$ and $p_Q - p_P$ are independent conditional on a single value of p_Q, meaning the conditional expectation of $(p_A - p_Q)(p_Q - p_P)$ is 0. By the double expectation theorem,

$$E((p_A - p_Q)(p_Q - p_P)) = E(E((p_A - p_Q)(p_Q - p_P)|p_Q)) = E(E(0)) = 0.$$

From $E(f_2^i(A, P)) = E(f_2^i(A, Q)) + E(f_2^i(Q, P))$, we can take the average over a set of SNPs to yield, in the notation from above,

$$F_2(A, P) = F_2(A, Q) + F_2(Q, P).$$

We have thus shown that f_2 distances are additive along an unadmixed-drift tree. This property is fundamental for our theoretical results and is also essential for finding admixtures, since, as we will see, additivity does not hold for admixed populations.

Given a set of populations with allele frequencies at a set of SNPs, we can use the estimator $\hat{f}_2$ to compute f_2 distances between each pair. These distances should be additive if the populations are related as a true tree. Thus, it is natural to build a phylogeny using neighbor-joining (Saitou and Nei, 1987), yielding a fully parameterized tree with all branch lengths inferred. However, in practice, the tree will not exactly be additive, and we may wish to try fitting some population C' as an admixture. To do so, we would have to specify six parameters (in the notation of Figure S6): the locations on the tree of A'' and B''; the branch lengths $f_2(A'', A)$, $f_2(B'', B)$, and $f_2(C, C')$; and the mixture fraction. These are the variables r, s, a, b, c, and α.

In order to fit C' onto an unadmixed tree (that is, solve for the six mixture parameters), we use the equations for the expectations $F_2(C', Z')$ of the f_2 distances between C' and each other population Z' in the tree. Referring to Figure S6, with the point admixture model, the allele

frequency in C is $p_C = \alpha\, p_A + (1-\alpha)\, p_B$. So, for a single locus, using additivity,

$$\begin{aligned}
E(f_2^i(A', C')) &= E((p_{A'} - p_{C'})^2) \\
&= E((p_{A'} - p_{A''} + p_{A''} - p_C + p_C - p_{C'})^2) \\
&= E((p_{A'} - p_{A''})^2) + E((p_{A''} - \alpha\, p_A - (1-\alpha)\, p_B)^2) + E((p_C - p_{C'})^2) \\
&= E(f_2^i(A', A'')) + \alpha^2 E(f_2^i(A'', A)) \\
&\quad +(1-\alpha)^2 E(f_2^i(A'', B)) + E(f_2^i(C, C')).
\end{aligned}$$

Averaging over SNPs, and replacing $E(f_2(A', C'))$ by the estimator $\hat{f}_2(A', C')$, this becomes

$$\begin{aligned}
\hat{f}_2(A', C') &= F_2(A', X'') - r + \alpha^2 a \\
&\quad +(1-\alpha)^2 (r + F_2(X'', Y'') + s + b) + c \\
\implies \hat{f}_2(A', C') - F_2(A', X'') &= (\alpha^2 - 2\alpha) r + (1-\alpha)^2 s + \alpha^2 a \\
&\quad +(1-\alpha)^2 b + c + (1-\alpha)^2 F_2(X'', Y'').
\end{aligned}$$

The quantities $F_2(X'', Y'')$ and $F_2(A', X'')$ are constants that can be read off of the neighbor-joining tree. Similarly, we have

$$\hat{f}_2(B', C') - F_2(B', Y'') = \alpha^2 r + (\alpha^2 - 1) s + \alpha^2 a + (1-\alpha)^2 b + c + \alpha^2 F_2(X'', Y'').$$

For the outgroups X' and Y', we have

$$\begin{aligned}
\hat{f}_2(X', C') &= \alpha^2 (c + a + r + F_2(X', X'')) \\
&\quad +(1-\alpha)^2 (c + b + s + F_2(X'', Y'') + F_2(X', X'')) \\
&\quad +2\alpha(1-\alpha)\, (c + F_2(X', X'')) \\
&= \alpha^2 r + (1-\alpha)^2 s + \alpha^2 a + (1-\alpha)^2 b + c \\
&\quad +(1-\alpha)^2 F_2(X'', Y'') + F_2(X', X'')
\end{aligned}$$

and

$$\hat{f}_2(Y', C') = \alpha^2 r + (1-\alpha)^2 s + \alpha^2 a + (1-\alpha)^2 b + c + \alpha^2 F_2(X'', Y'') + F_2(Y', Y'').$$

Assuming additivity within the neighbor-joining tree, any population descended from A'' will give the same equation (the first type), as will any population descended from B'' (the second type), and any outgroup (the third type, up to a constant and a coefficient of α). Thus, no matter how many populations there are in the unadmixed tree—and assuming there are at least two outgroups X' and Y' such that the points X'' and Y'' are distinct—the system of equations consisting of $E(f_2(P, C'))$ for all P will contain precisely enough information to solve for α, r, s, and the linear combination $\alpha^2 a + (1-\alpha)^2 b + c$. We also note the useful fact that for a fixed value of α, the system is linear in the remaining variables.

Text S2. Heterozygosity and drift lengths.

One disadvantage to building trees with f_2 statistics is that the values are not in easily interpretable units. For a single locus, the f_2 statistic measures the squared allele frequency change between two populations. However, in practice, one needs to compute an average f_2 value over many loci. Since the amount of drift per generation is proportional to $p(1-p)$, the expected frequency change in a given time interval will be different for loci with different initial frequencies. This means that the estimator $\hat{f}_2$ depends on the distribution of frequencies of the SNPs used to calculate it. For example, within an f_2-based phylogeny, the lengths of non-adjacent edges are not directly comparable.

In order to make use of the properties of f_2 statistics for admixture tree building and still be able to present our final trees in more directly meaningful units, we will show now how f_2 distances can be converted into absolute drift lengths. Again, we consider a biallelic, neutral SNP in two populations, with no further mutations, under a Wright-Fisher model of genetic drift.

Suppose populations A and B are descended independently from a population P, and we have an allele with frequency p in P, $p_A = p + a$ in A, and $p_B = p + b$ in B. The (true) heterozygosities at this locus are $h^i_P = 2p(1-p)$, $h^i_A = 2p_A(1-p_A)$, and $h^i_B = 2p_B(1-p_B)$. As above, we write $\hat{h}^i_A$ for the unbiased single-locus estimator

$$\hat{h}^i_A := \frac{2n_A\hat{p}_A(1-\hat{p}_A)}{n_A - 1},$$

$\hat{h}_A$ for the multi-locus average of $\hat{h}^i_A$, and H^i_A for the expectation of h^i_A under the Wright-Fisher model (and similarly for B and P).

Say A has experienced t_A generations of drift with effective population size N_A since the split from P, and B has experienced t_B generations of drift with effective population size N_B. Then it is well known that $H^i_A = h^i_P(1 - D_A)$, where $D_A = 1 - (1 - 1/(2N_A))^{t_A}$, and

$H_B^i = h_P^i(1 - D_B)$. We also have

$$\begin{aligned} H_A^i &= E(2(p+a)(1-p-a)) \\ &= E(h_P^i - 2ap + 2a - 2ap - 2a^2) \\ &= h_P^i - 2E(a^2) \\ &= h_P^i - 2F_2^i(A,P), \end{aligned}$$

so $2F_2^i(A,P) = h_P^i D_A$. Likewise, $2F_2^i(B,P) = h_P^i D_B$ and $2F_2^i(A,B) = h_P^i(D_A + D_B)$. Finally,

$$H_A^i + H_B^i + 2F_2^i(A,B) = h_P^i(1 - D_A) + h_P^i(1 - D_B) + h_P^i(D_A + D_B) = 2h_P^i.$$

This equation is essentially equivalent to one in Nei (1987), although Nei interprets his version as a way to calculate the expected present-day heterozygosity rather than estimate the ancestral heterozygosity. To our knowledge, the equation has not been applied in the past for this second purpose.

In terms of allele frequencies, the form of h_P^i turns out to be very simple:

$$h_P^i = p_A + p_B - 2p_A p_B = p_A(1 - p_B) + p_B(1 - p_A),$$

which is the probability that two alleles, one sampled from A and one from B, are different by state. We can see, therefore, that this probability remains constant in expectation after any amount of drift in A and B. This fact is easily proved directly:

$$E(p_A + p_B - 2p_A p_B) = 2p - 2p^2 = h_P^i,$$

where we use the independence of drift in A and B.

Let $\hat{h}_P^i := (\hat{h}_A^i + \hat{h}_B^i + 2\hat{f}_2^i(A,B))/2$, and let h_P denote the true average heterozygosity in P

over an entire set of SNPs. Since $\hat{h}_P^i$ is an unbiased estimator of $(h_A^i + h_B^i + 2f_2^i(A,B))/2$, its expectation under the Wright-Fisher model is h_P^i. So, the average $\hat{h}_P$ of $\hat{h}_P^i$ over a set of SNPs is an unbiased (and potentially low-variance) estimator of h_P. If we have already constructed a phylogenetic tree using pairwise f_2 statistics, we can use the inferred branch length $\hat{f}_2(A', P)$ from a present-day population A to an ancestor P in order to estimate $\hat{h}_P$ more directly as $\hat{h}_P = \hat{h}_A + 2\hat{f}_2(A,P)$. This allows us, for example, to estimate heterozygosities at intermediate points along branches or in the ancestors of present-day admixed populations.

The statistic $\hat{h}_P$ is interesting in its own right, as it gives an unbiased estimate of the heterozygosity in the common ancestor of any pair of populations (for a certain subset of the genome). For our purposes, though, it is most useful because we can form the quotient

$$\hat{d}_A := \frac{2\hat{f}_2(A,P)}{\hat{h}_P},$$

where the f_2 statistic is inferred from a tree. This statistic $\hat{d}_A$ is not exactly unbiased, but by the law of large numbers, if we use many SNPs, its expectation is very nearly

$$E(\hat{d}_A) \approx \frac{E(2\hat{f}_2(A,P))}{E(\hat{h}_P)} = \frac{h_P D_A}{h_P} = D_A,$$

where we use the fact that D_A is the same for all loci. Thus $\hat{d}$ is a simple, direct, nearly unbiased moment estimator for the drift length between a population and one of its ancestors. This allows us to convert branch lengths from f_2 distances into absolute drift lengths, one branch at a time, by inferring ancestral heterozygosities and then dividing.

For a terminal admixed branch leading to a present-day population C' with heterozygosity $\hat{h}_{C'}$, we divide twice the inferred mixed drift $c_1 = \alpha^2 a + (1-\alpha)^2 b + c$ (Figure 2) by the heterozygosity $\hat{h}^*_{C'} := \hat{h}_{C'} + 2c_1$. This is only an approximate conversion, since it utilizes a common value $\hat{h}^*_{C'}$ for what are really three disjoint branches, but the error should be very small with short drifts.

An alternative definition of $\hat{d}_A$ would be $1 - \hat{h}_A/\hat{h}_P$, which also has expectation (roughly)

D_A. In most cases, we prefer to use the definition in the previous paragraph, which allows us to leverage the greater robustness of the f_2 statistics, especially when taken from a multi-population tree.

We note that this estimate of drift lengths is similar in spirit to the widely-used statistic F_{ST}. For example, under proper conditions, the expectation of F_{ST} among populations that have diverged under unadmixed drift is also $1-(1-1/(2N_e))^t$ (Nei, 1987). When F_{ST} is calculated for two populations at a biallelic locus using the formula $(\Pi_D - \Pi_S)/\Pi_D$, where Π_D is the probability two alleles from different populations are different by state and Π_S is the (average) probability two alleles from the same population are different by state (as in Reich et al. (2009) or the measure G'_{ST} in Nei (1987)), then this F_{ST} is exactly half of our $\hat{d}$. As a general rule, drift lengths $\hat{d}$ are approximately twice as large as values of F_{ST} reported elsewhere.