Supplementary Material: Fine-scale Inference of Ancestry Segments without Prior Knowledge of Admixing Groups

Michael Salter-Townshend and Simon Myers

S1 Comparison with GLOBETROTTER for 2-way admixture events

As we regard GLOBETROTTER to be the closest in spirit to our approach (in terms of genome-wide estimate), we expand upon the comparison provided in Section Two-way, Single Event of main paper. Table S1 compares results for the two-way admixture events inferred from the expanded HGDP dataset and comprises the values used to create Figure 3 in the main paper. The contribution of our method lies in accurate multiway local ancestry estimation within a single model and framework that provides these estimates.

Table S2 shows the number of individuals in each population in the extended HGDP dataset analyzed in the main text Section Application to Human Genome Diversity Project Data.

S2 Bootstrapped Dates for Coancestry Curves

S2.1 Simulated Data

We performed Bootstrapping (see Section Extended Human Genome Diversity Project Two-way, Single Event of main paper) of the chromosomal local ancestries for the simulation example used in Figure 1b and Figure 1c of the main paper. In the simulation, the same admixture date of 50 generations was used on both simulated individuals. As per the main paper, we forced the coancestry exponential decay curve fitting to use a single scalar λ to reflect an assumption that the decay curves share a single date (Figure S1 black line). We show the same bootstrap analysis with that assumption relaxed so that the decay of a:a, a:b, and b:b switches each depend on a different decay rate (Figure S1 blue line).

When multiway admixture models are fit the interpretation of the pairwise coancestry curve decay parameters is not straightforward (see Section Interpretation of Pairwise Decay Parameters for Multiway Admixture of main paper).

Population	n	source 1	source 2	Rst	GLOBETROTTER	MOSAIC
Hazara	22	Pathan	Mongola	0.0931	22 ± 0.9	20 ± 0.7
Uzbekistani	15	Turkish	Mongola	0.102	19 ± 1.1	19 ± 0.8
Uygur	10	Iranian	Mongola	0.101	22 ± 1.3	22 ± 1
Makrani	22	Balochi	BantuKenya	0.124	18 ± 1.2	16 ± 0.9
Druze	42	Cypriot	Ethiopian	0.159	37 ± 1.9	35 ± 2.1
Mozabite	25	Moroccan	Yoruba	0.122	21 ± 1.3	28 ± 1.4
Turkish	17	Armenian	Uygur	0.0662	24 ± 1.5	22 ± 1.1
Brahui	23	Balochi	Ethiopian	0.0885	20 ± 1.5	16 ± 2.2
Yemeni	4	Jordanian	Ethiopian	0.0954	14 ± 1.8	15 ± 1.8
Pima	14	Egyptian	Maya	0.0422	6 ± 0.9	6 ± 1
BantuSA	8	SanKhomani	Yoruba	0.0113	25 ± 2.3	24 ± 1.8
Tu	10	Turkish	HanNchina	0.0947	25 ± 2.3	23 ± 1.2
W.Sicilian	10	E.Sicilian	Ethiopian	0.117	27 ± 3.9	25 ± 4.6
Cambodian	10	Uygur	Dai	0.035	20 ± 2.7	17 ± 0.8
Georgian	20	Armenian	Russian	0.00338	30 ± 3.3	8 ± 1.1
Romanian	13	Bulgarian	Uygur	0.0545	31 ± 2.6	23 ± 2.3
Bulgarian	18	Romanian	Uzbekistani	0.0433	28 ± 3.5	25 ± 1.7
Hezhen	8	Uzbekistani	Daur	0.0711	13 ± 1.3	22 ± 2.2
Oroqen	9	Uzbekistani	Daur	0.0855	15 ± 2	22 ± 2.4
Hungarian	18	GerAus	Uygur	0.0535	39 ± 3.5	35 ± 1.5
HanNchina	10	Uzbekistani	Tujia	0.082	26 ± 3.8	29 ± 1.4
Daur	9	Turkish	Mongola	0.0814	21 ± 1.7	19 ± 1.3
Greek	20	E.Sicilian	Ethiopian	0.0493	36 ± 3.7	37 ± 2.6
Melanesian	10	Sandawe	Papuan	0.0475	28 ± 7.6	221 ± 41.7
Mandenka	22	Yoruba	Ethiopian	0.0438	19 ± 4.2	21 ± 1.4
Indian	13	Sindhi	Sindhi	0.00332	53 ± 8.4	1 ± 0.3
N.Italian	12	Spanish	Tunisian	0.0694	71 ± 11.8	30 ± 6.8
Polish	16	Belorussian	Uzbekistani	0.0433	31 ± 5.1	28 ± 2.8
Tuscan	8	W.Sicilian	Ethiopian	0.0954	35 ± 6.1	37 ± 3.1
SanNamibia	5	SanKhomani	SanKhomani	0.0187	48 ± 8.9	16 ± 1

Table S1: Comparison of date estimates from MOSAIC and from GLOBETROTTER for all inferred 2-way admixed populations in the extended HGDP dataset

Adygei	17	Armenian	16	Balochi	21	BantuKenya	11	BantuSA	8
Basque	24	Bedouin	45	Belorussian	8	BiakaPygmy	21	Brahui	23
Bulgarian	18	Burusho	25	Cambodian	10	Chuvash	16	Colombian	7
Cypriot	12	Dai	10	Daur	9	Druze	42	E.Sicilian	10
Egyptian	10	English	6	Ethiopian	19	EthiopianJew	11	Finnish	2
French	28	Georgian	20	GerAus	4	Greek	20	Hadza	3
Han	34	HanNchina	10	Hazara	22	Hezhen	8	Hungarian	18
Indian	13	IndianJew	8	Iranian	13	Ireland	7	Japanese	28
Jordanian	18	Kalash	23	Karitiana	14	Lahu	8	Lezgin	18
Lithuanian	10	Makrani	22	Mandenka	22	Maya	21	MbutiPygmy	13
Melanesian	10	Miao	10	Mongola	10	Moroccan	22	Mozabite	25
Myanmar	3	Naxi	8	N.Italian	12	Norwegian	18	Orcadian	15
Oroqen	9	Palestinian	46	Papuan	16	Pathan	22	Pima	14
Polish	16	Romanian	13	Russian	25	Sandawe	28	SanKhomani	30
SanNamibia	5	Sardinian	28	Saudi	10	Scottish	6	She	10
Sindhi	23	S.Italian	18	Spanish	34	Surui	8	Syrian	16
Tu	10	Tujia	10	Tunisian	12	Turkish	17	Tuscan	8
UAE	9	Uygur	10	Uzbekistani	15	Welsh	4	W.Sicilian	10
Xibo	9	Yakut	25	Yemeni	4	Yi	10	Yoruba	21

Table S2: Number of samples in each population analysed. Additional details are available in Table S.10 of Hellenthal *et al.* (2014), including original source study for each population.

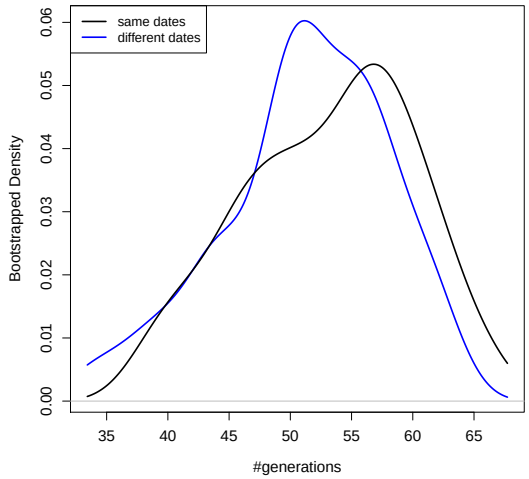
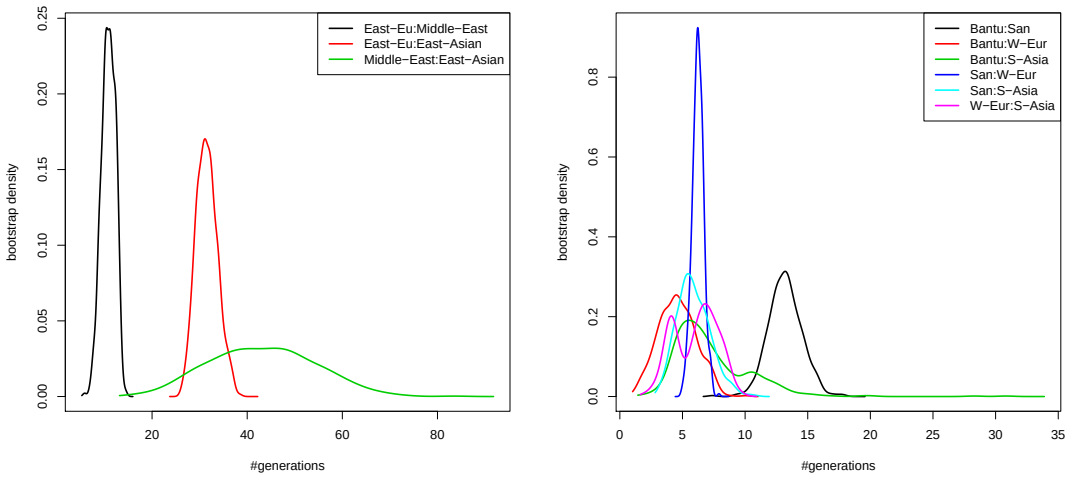


Figure S1: True admixure date of 50 generations ago is well captured by the inferred dates. Although Figure 1c of the main text restricts the decay curve to a single estimated λ shared across curves, this assumption is relaxed for subsequent analysis.

S2.2 Chuvash 3-way and San-Khomani 4-way Admixture

Figure S2 shows the bootstrapped estimated pairwise dates of admixture for Chuvash 3-way and San-Khomani 4-way models which are used to infer order and timings of admixture events as pairwise meetings of each ancestral component.



(a) Chuvash 3-way.

(b) San-Khomani 4-way.

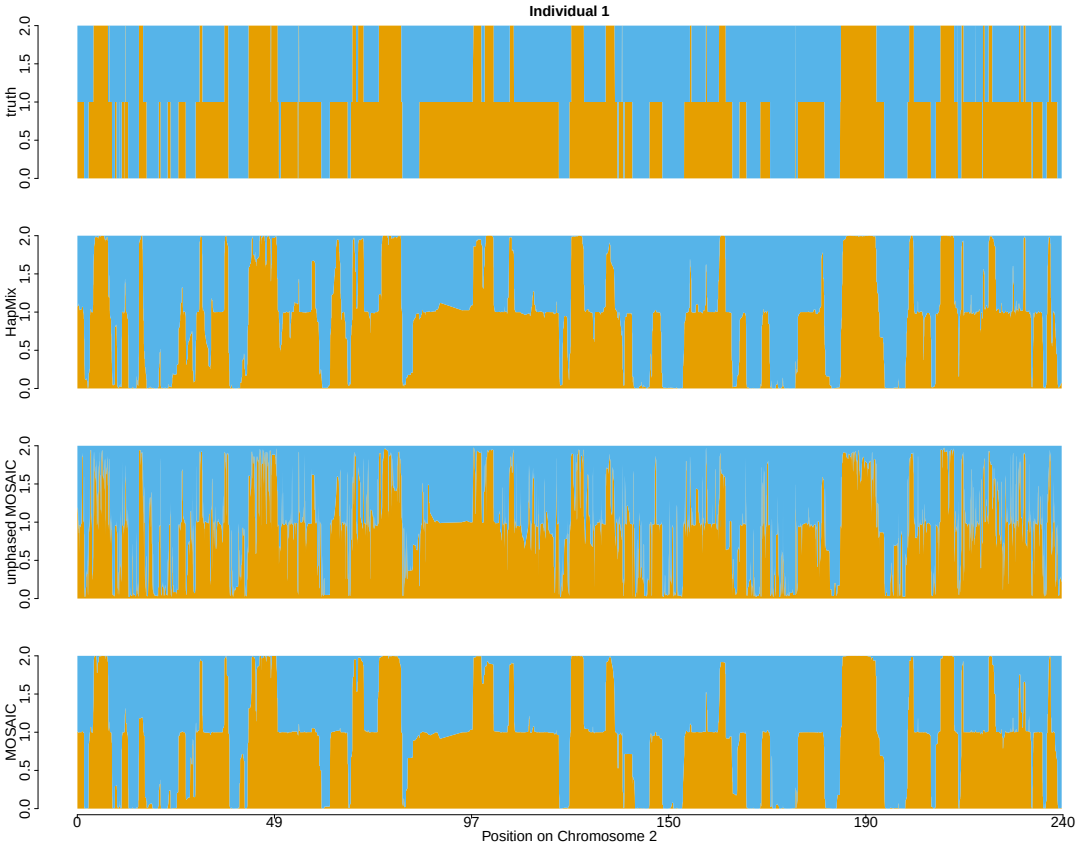
Figure S2: Bootstrapped estimated pairwise generations since mixing for admixture events inferred from Chuvash and San-Khomani samples.

S3 Phasing: Simulation Example

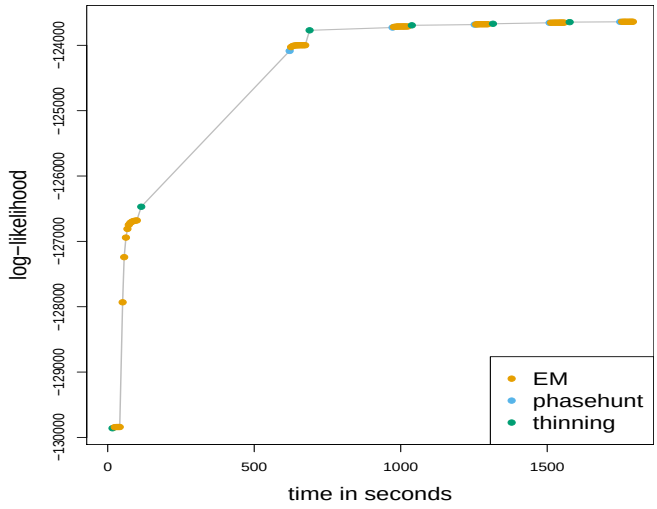
Repeating the simulation example of Figure 2 (easy simulation in Section Simulation Studies) of the main paper is presented in Figure S3a below, with the additional detail of MOSAIC result without ancestry aware re-phasing and log-likelihood trace plot across the MOSAIC inferential algorithm. This demonstrates the utility of the phase hunting method within MOSAIC. Figure S3b shows the progress of the algorithm as increasing log-likelihood against time.

S4 2-way Simulation Results without Direct Surrogates

We present results here for simulated admixture followed by MOSAIC inference using panels that do not correspond to unmixed surrogates for the mixing groups. Specifically, we create admixed chromosomes 1 and 2 by recombining Spanish and Yoruban chromosomes from 8 individuals with ancestral recombination segment



(a) Example diploid local ancestry in simulated dataset. From top down: true local ancestry, HapMix inferred local ancestry, MOSAIC inferred local ancestry without ancestry informed re-phasing, MOSAIC inferred local ancestry with re-phasing.



(b) Trace plot of log-likelihood over time for MOSAIC inference of simulated example.

Figure S3: Comparison of local ancestry estimation using MOSAIC and HapMix in 2-way setting on simulated data, with relatively accurate proxy reference panels (top) and algorithm progress (bottom). The improvement due to the ancestry sensitive re-phasing can be seen clearly.

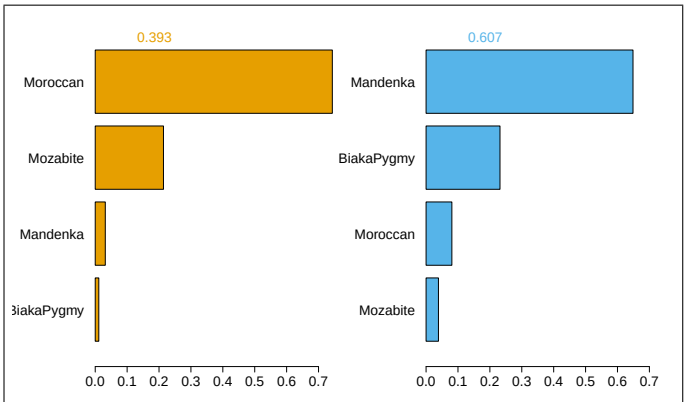
lengths based on 50 generations since admixture. We then run MOSAIC using only Moroccan, Mozabite, Mandenka, and Biaka-Pygmy reference haplotypes (without un-admixed Spanish-like reference panels). MOSAIC infers all parameters and local ancestry estimates. The copying matrix depicted in Figure S4a shows that Moroccan (and to a lesser extent Mozabite) haplotypes are copied by both sides, reflecting the fact that these reference panels contain both European and African ancestry. When $\hat{F}_{\text{st}}$ is calculated it demonstrates that none of the reference panels are extremely close in F_{st} to the first (Spanish) ancestry, moreover there is a very high degree of correlation between the inferred $\hat{F}_{\text{st}}$ from each panel to each latent ancestry (computed via the partial reconstructed genomes) and the $\hat{F}_{\text{st}}$ between the panels and the Spanish and Yoruban data used to simulate the admixture (Pearson sample correlation of 0.998) as shown in Table S3. In this example, the MOSAIC inferred $\hat{F}_{\text{st}}$ between the mixing groups is 0.15 and the $\hat{F}_{\text{st}}$ between the Spanish and Yoruban panels is 0.147. Note here that by necessity not all of the chromosomes in the Spanish and Yoruban samples can be used to simulate admixture so that some discrepancy is expected even for perfect inference of local ancestry. Indeed, the $\hat{F}_{\text{st}}$ between partial genomes created using the **known true** local ancestry for the two groups is 0.156 here. Finally, accurate genome-wide ancestry proportions for each individual were obtained with a correlation of 0.996 between true and inferred proportions of ancestry type 1 in each individual where the simulated range is for type 1 is between 0.396 and 0.782. The coancestry curves are shown in Figure S4c and demonstrate MOSAIC's accurate date estimation, even when the reference panels are imperfect. Note here that $r^2 = 0.791$ with the true local ancestry.

	Spanish	Yoruba
Moroccan	0.010 (0.012)	0.104 (0.11)
Mozabite	0.025 (0.026)	0.106 (0.11)
Mandenka	0.151 (0.145)	0.007 (0.009)
BiakaPygmy	0.184 (0.172)	0.044 (0.041)

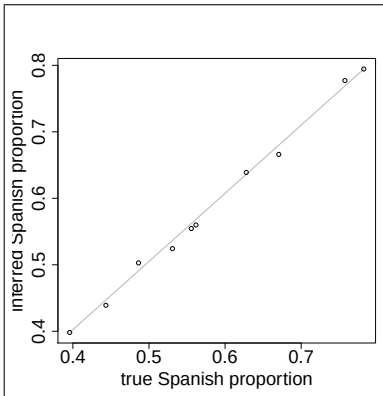
Table S3: Estimated via partial genomes and true (in brackets) $\hat{F}_{\text{st}}$ between haplotype panels used to simulate admixture (Spanish and Yoruba) and those used by MOSAIC to infer admixture. MOSAIC can accurately infer genetic differentiation between extant panels and unseen mixing groups.

S5 Comparison with LAMP-LD, RFMix, and ELAI

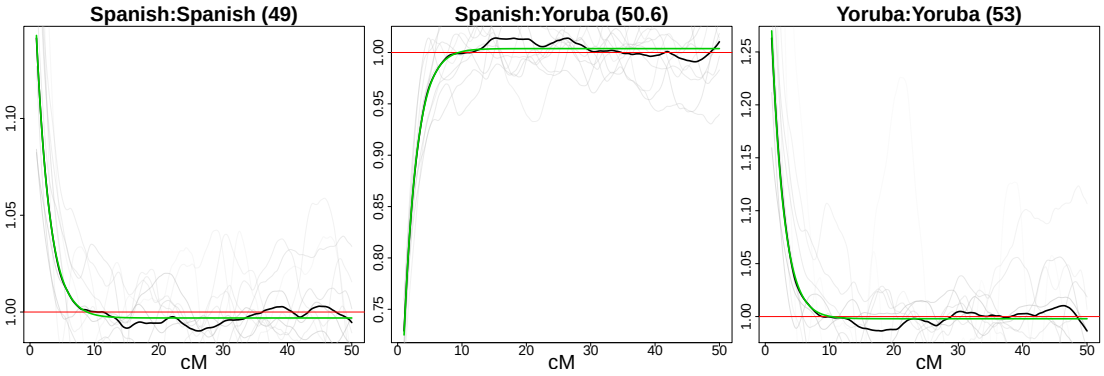
LAMP-LD (Baran *et al.* 2012), RFMix (Maples *et al.* 2013) and ELAI (Guan 2014) are among the leading existing approaches to infer local ancestry from multi-way admixed samples. See the main text for brief descriptions of these methods. In



(a) Inferred copying matrix $\boldsymbol{\mu}$ as bar plots. Figure S7 uses colour density to allow a greater number of results to be shown on the same plot.



(b) Inferred versus true genome-wide ancestry proportions for each individual. The use of imperfect reference panels does not appear to bias ancestry estimation.



(c) Coancestry curves from inferred ancestry segments.

Figure S4: Results of MOSAIC on admixture simulated between Spanish and Yoruban individuals 50 generations ago, on chromosomes 1 and 2.

the below we compare accuracy of local ancestry inference across these methods in a 3-way admixture simulation.

3-way admixture 50 generations ago was simulated in four individuals using haplotypes from Europe, Africa, and Asia 50 generations ago in equal proportions in Chromosome 1 only. MOSAIC, LAMP-LD, ELAI, and RFMix v1.5.4 were then applied to infer local ancestry tracts along this chromosome. Panels were provided coming from the same continent as the mixing populations; French, Mandenka, and Han. MOSAIC infers the stochastic relationships between the ancestral groups and the donor panels which are

- **Europe:** English, GerAus, Spanish.
- **Africa:** Yoruba, BiakaPygmy, Sandawe.
- **Asia:** Daur, Mongola, and Oroqen.

LAMP-LD, ELAI, and RFMix must all be provided with known amalgamations of these into respective continents and the parameters of all three were optimised to maximise the correlation with the truth (15 ancestral founder haplotypes; window size of 50 for LAMP-LD; ELAI and RFMix were run using the known generations since mixture of 50; ELAI was run with 5 lower clusters per upper cluster; RFMix window size was set to $1/50cM$; ELAI was run in diploid mode). ELAI and RFMix were both run in EM mode to further improve model fit and the RFMix forward backward probabilities were output. This represents an "easy scenario" as the reference panels for the mixing groups are not markedly admixed with respect to the other groups and each of the three ancestries is from a different continent with a high degree of drift between them.

Following adaptation of input files to the format required by each method, the following commands were used:

MOSAIC:

```
Rscript mosaic.R simulated MOSAIC/inputs/ -a 3 -c 1:1 -n 4
        -p "French Mandenka Han
            English GerAus Spanish
            Yoruba BiakaPygmy Sandawe
            Daur Mongola Oroqen"
```

LAMP-LD:

```
unolanc 15 15 chr1.pos European_1haps.ref African_1haps.ref Asian_1haps.ref
                          admixed_3way_1.gen lampped_3way_1.out.lanc
perl convertLAMPLDout.pl lampped_3way_1.out.lanc lampped_3way_1.out.long
```

ELAI:

```
./elai-lin -g European_1.inp -p 10 -g African_1.inp -p 11 -g Asian_1.inp -p 12 -C 3 -c 15
           -g simulated_3way_1.inp -p 1 -pos snp.chr1.pos -o simulated_3way_1 -s 30 -mg 50
```

RFMix:

```
python2.7 RunRFMix.py PopPhased alleles_3way_1.txt classes_3way_1.txt map.1 -e 5 -w 0.02
                      -G 50 -o simulated_3way_1 --forward-backward
```

We then repeated this "easy scenario" experiment for 5, 10, 20, 50, and 100 generations since mixing, again adjusting the parameters for all methods to maximize correlation with the true local ancestry.

To explore the impact of imperfect reference panels we then repeated the same experiment with the European panels replaced by Moroccan, Mozabite, and Tunisian donors ("hard scenario"). Although these represent a better proxy for the European mixing group (French) than the other potential donors from African and Asia, they are themselves an admixture of Sub-Saharan African and European populations (see for example Figure S7 and Table S1). This is similar to the experiment for 2-way admixture described in Section S4. All four of the methods cope surprisingly well, but again MOSAIC is able to outperform the others. The squared correlation r^2 between the true diploid local ancestry and inferred local ancestry for each method is shown in Table S4 and the local ancestry for a single individual is shown in Figure S5 for simulations again involving 5, 10, 20, 50, and 100 generations since a single admixture event.

r^2 of local ancestry with true simulated values.

European	#Gens	LAMP-LD	ELAI	RFMix	MOSAIC
EU	5	0.918	0.929	0.958	**0.967**
EU	10	0.885	0.893	0.933	**0.946**
EU	20	0.835	0.828	0.892	**0.910**
EU	50	0.722	0.732	0.786	**0.819**
EU	100	0.541	0.645	0.642	**0.711**
NA	5	0.851	0.860	0.930	**0.951**
NA	10	0.786	0.806	0.886	**0.935**
NA	20	0.709	0.766	0.850	**0.869**
NA	50	0.566	0.655	0.709	**0.775**
NA	100	0.396	0.521	0.585	**0.623**

Table S4: r^2 between true (simulated) 3-way local ancestry and estimated local ancestry using LAMP-LD, ELAI, RFMix, and MOSAIC. The first column shows whether admixed North African (NA, "hard scenario") or European (EU, "easy scenario") surrogates are used for the French ancestry. All methods except MOSAIC require knowledge of which donors relate to which ancestry whereas MOSAIC infers the relationships.

For the above simulation study, when averaged over 5 replications the run times in seconds for each method run on 4 individuals are: MOSAIC (263); LAMP-LD (442); RFMix (91); ELAI (624).

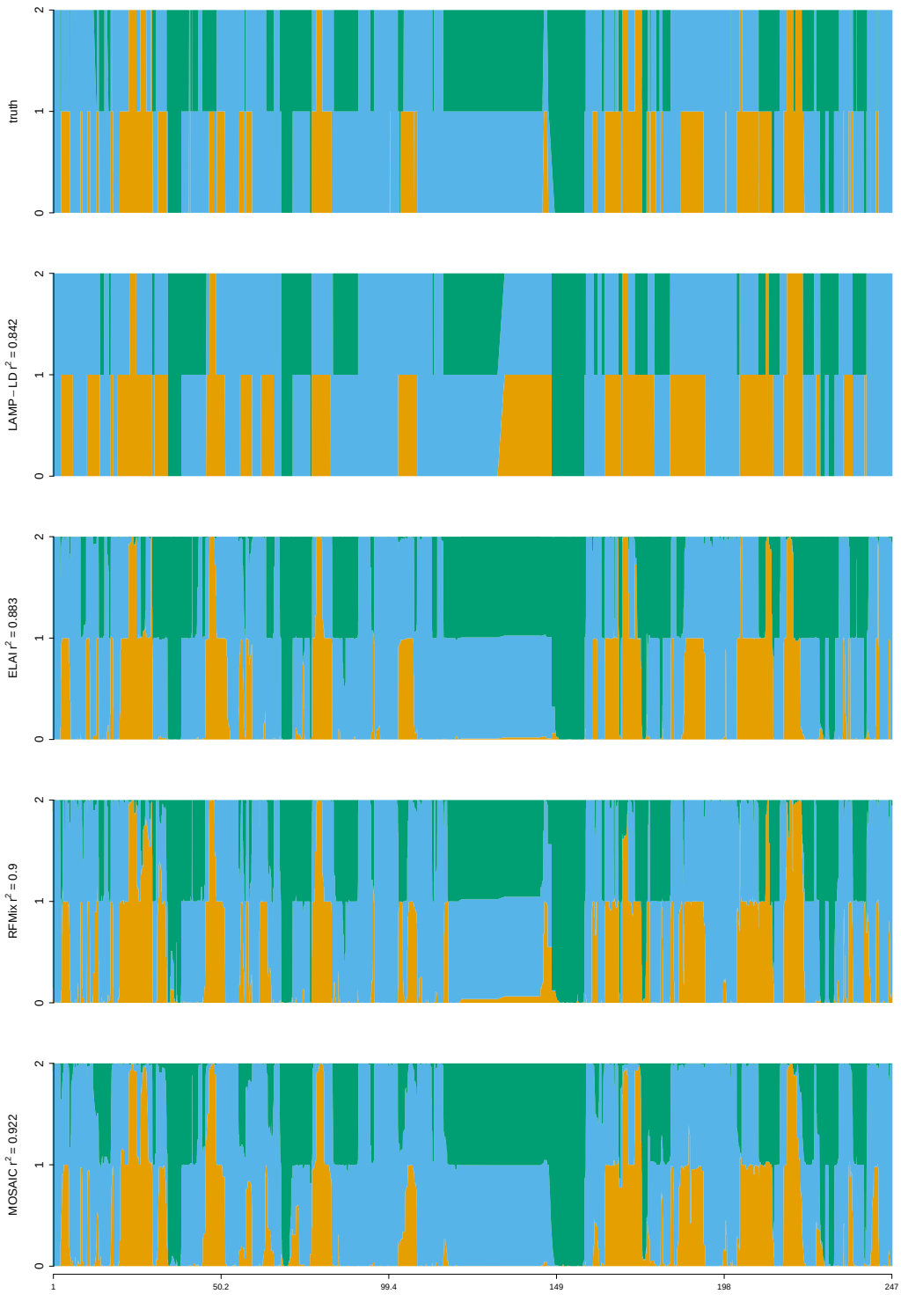


Figure S5: Comparison between LAMP-LD, ELAI, RFMix, and MOSAIC on a simulated 3-way admixture event with imperfect reference panels. Local ancestry along one chromosome for one individual is shown. Each y-axis gives the method and the squared correlation to the truth along just this chromosome for this individual.

	French	Mandeka	Han
Moroccan	0.0112	0.0999	0.0811
Mozabite	0.0255	0.1017	0.0953
Tunisian	0.0094	0.1000	0.0820
Yoruba	0.1597	0.0059	0.1790
BiakaPygmy	0.2008	0.0527	0.2195
Sandawe	0.1137	0.0277	0.1406
Daur	0.1305	0.1938	0.0067
Mongola	0.1144	0.1811	0.0019
Oroqen	0.1267	0.1932	0.0193

Table S5: Estimated via partial genomes $\hat{F}_{\text{st}}$ between haplotype panels used to simulate admixture (French, Mandenka, and Han) and those used by MOSAIC to infer admixture. This is "hard scenario" with 20 generations since 3-way admixture (corresponding to row 8 of Table S4). MOSAIC obtains estimates of local ancestry with $r^2 = 0.869$ to the true local ancestry.

S5.1 Admixed Panels

To further explore the impact of imperfect panels, we ran a simulated data experiment to assess the impact of the inclusion of reference panels with the *same* admixture profile as the target admixed individuals. We tested the robustness of MOSAIC to this scenario by creating 8 admixed individuals from French, Mandenka, and Han chromosome 1s 50 generations ago. We then placed 6 of these in a new reference panel and made the other 2 the target individuals. We ran MOSAIC using this admixed panel along with English, Germany-Austria, Spanish, Yoruba, Biaka-Pygmy, Sandawe, Daur, Mongola, and Oroqen panels (i.e. similar to the above "easy scenario").

Figure S6a shows the resulting inferred copying matrix $\boldsymbol{\mu}$. In this scenario, MOSAIC is still able to achieve $r^2 = 0.926$ despite the presence of the admixed panel that is composed of genome segments from the same original HGDP panels as the target genomes. Although the admixed panel (labelled French/Mandenka/Han) is copied from extensively by segments of all 3 ancestries in the target chromosomes this does not appear to impact on MOSAIC's ability to infer accurate local ancestry. The true ancestry proportions across the simulated targets is $0.266, 0.434, 0.3$ respectively. Note that LAMP-LD, RFMix, and ELAI cannot be used in this scenario as the admixed panel cannot be placed into just one set of references.

When we reduce the available panels further, MOSAIC does of course lose accuracy of local ancestry estimation. Figure S6b shows the copying matrix inferred when provided with a single putatively un-admixed (with respect to the target admixture) reference panel for each ancestry (Figure S6b). As always, MOSAIC infers the stochastic relationships between ancestries (created from French, Mandenka, Han) and the 4 panels (English, Yoruba, Daur, and French/Mandenka/Han). Local ancestry estimation accuracy is $r^2 = 0.62$ due to

the reduced reference haplotypes upon which to learn the model parameters and estimate local ancestry, however MOSAIC does not appear to be overly biased towards copying from the simulated admixed panel.

Finally, if MOSAIC is provided with no good surrogates for one ancestry (French) then copying for this ancestry is highest to the admixed panel (Figure S6c). Now accuracy is degraded to $r^2 = 0.47$, with ability to infer the French segments particularly impacted; $r^2 = 0.6$ when local ancestry between the 2nd and 3rd ancestries only are examined.

S6 Additional Case Studies

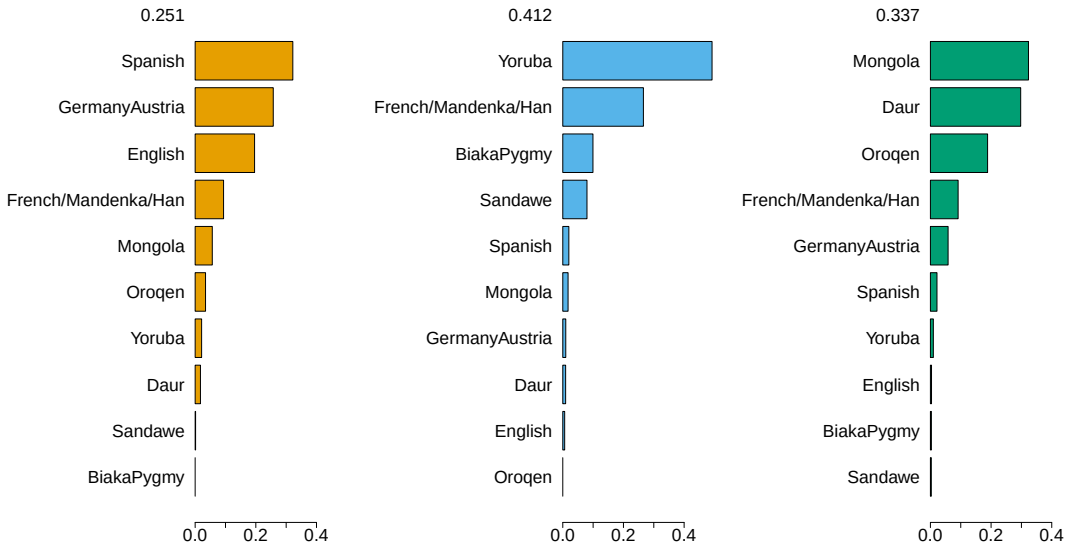
We present some additional case studies of MOSAIC applied to the extended HGDP dataset from `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53626`.

S6.1 Moroccan 2-way Admixture

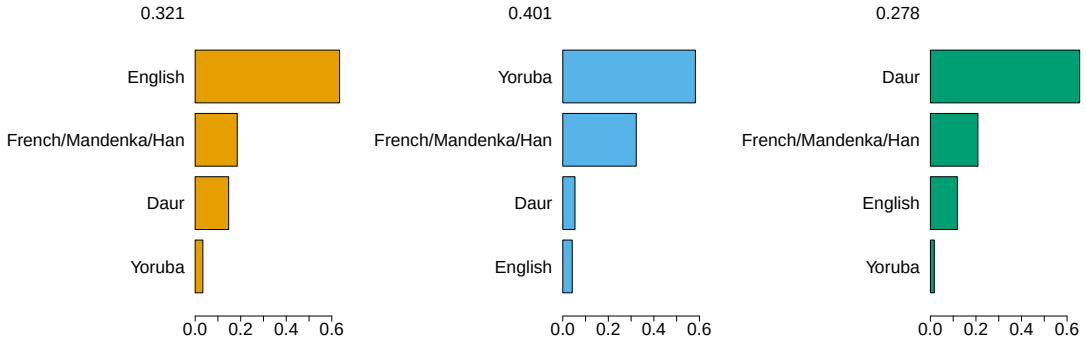
Results of fitting a 2-way admixture model to Moroccan people in North Africa, with masking (not allowing copying from) of local (North African) populations. Similar results are obtained when these local groups are included (see online at `https://maths.ucd.ie/~mst/MOSAIC/HGDP_browser/Moroccan` to compare both), with the masked version obtaining a somewhat older estimated date (37.5 versus 32 generations ago). Similarly to the 2-way admixture analysis of the Bedouin described in the main paper, the major ancestral group copies from populations close to but outside African (Southern European) and the minor (17%) ancestral component is related to modern sub-Saharan populations. See Table S6 for the 5 closest modern populations in terms of F_{st} to the inferred mixing groups. Figure S7 shows the copying proportions for the two ancestries and Figure S8a demonstrates a good fit to a single admixture event as modelled by coancestry curves.

S.Italian	0.006	Mandenka	0.023
W.Sicilian	0.0061	Yoruba	0.023
E.Sicilian	0.0061	BantuKenya	0.025
Spanish	0.0068	BantuSA	0.03
Greek	0.0073	Sandawe	0.037

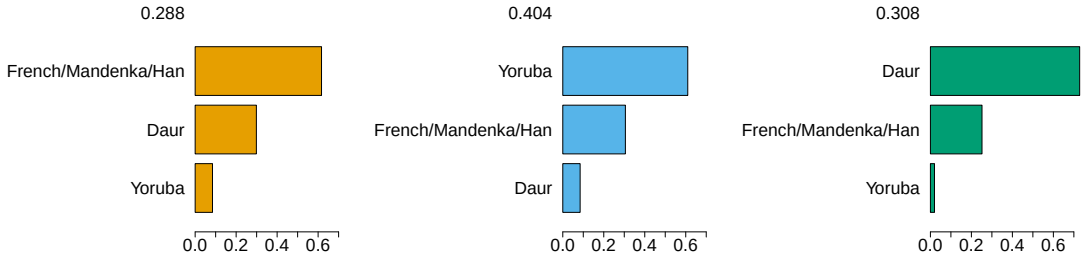
Table S6: F_{st} estimates between local ancestries and the closest 5 panels in fit of a 2-way admixture event in Moroccans. The F_{st} estimate between the inferred local ancestries is 1x2=0.14. The R_{st} is 0.15.



(a) Presence of the admixed panel within 10 reference panels.



(b) Presence of the admixed panel within 4 reference panels.



(c) Presence of the admixed panel within 3 reference panels. The French ancestry now mostly copies from the admixed panel.

Figure S6: Inferred copying matrices $\boldsymbol{\mu}$ for simulations including an admixed reference panel in Section S5.1. Along the top are the marginal ancestry proportions. Although MOSAIC infers substantial copying from the admixed reference panel (labelled French/Mandenka/Han) across all 3 ancestries, it is not the most copied from panel for any of the 3 ancestries despite the target chromosomes being made up of other individuals from the same 3 original panels unless no other panel contains unadmixed donors (lowest figure).

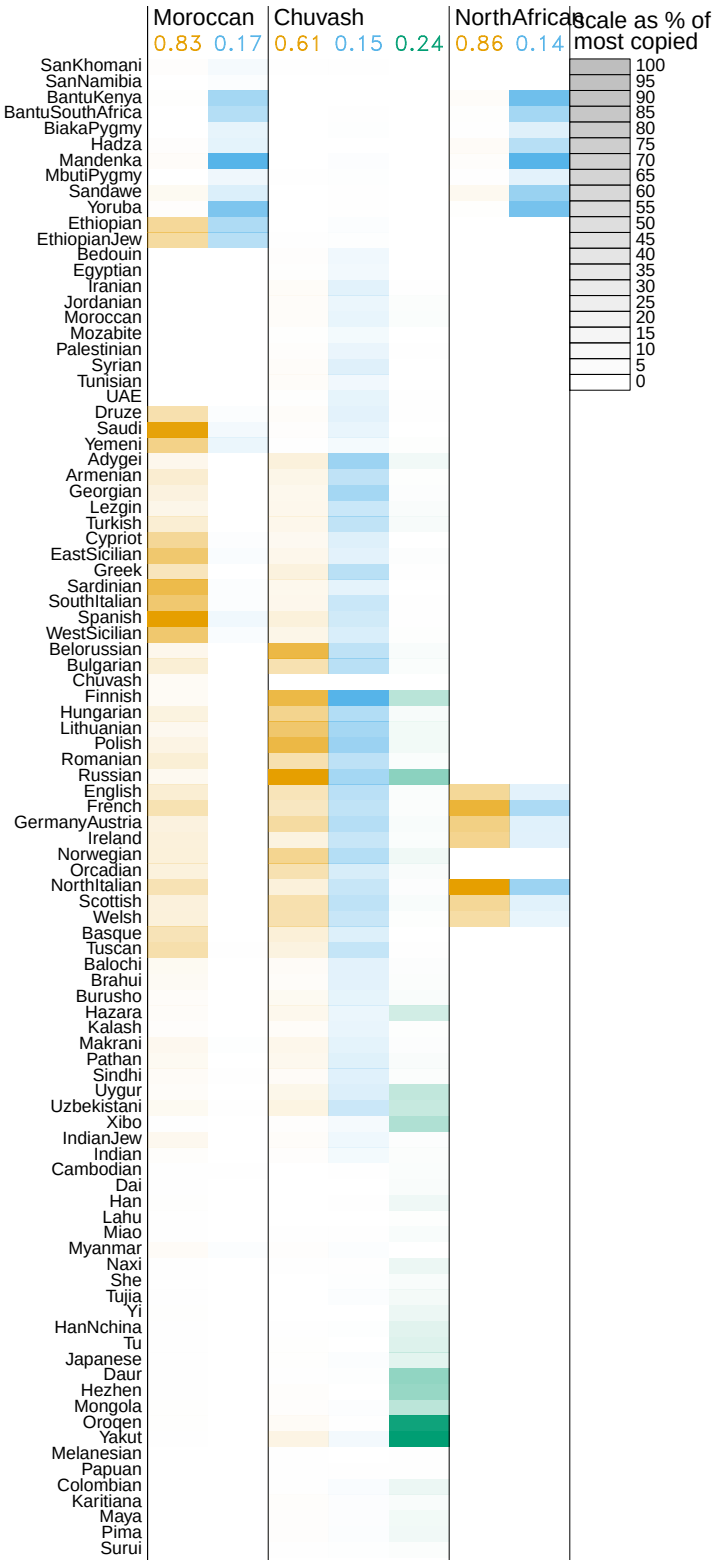
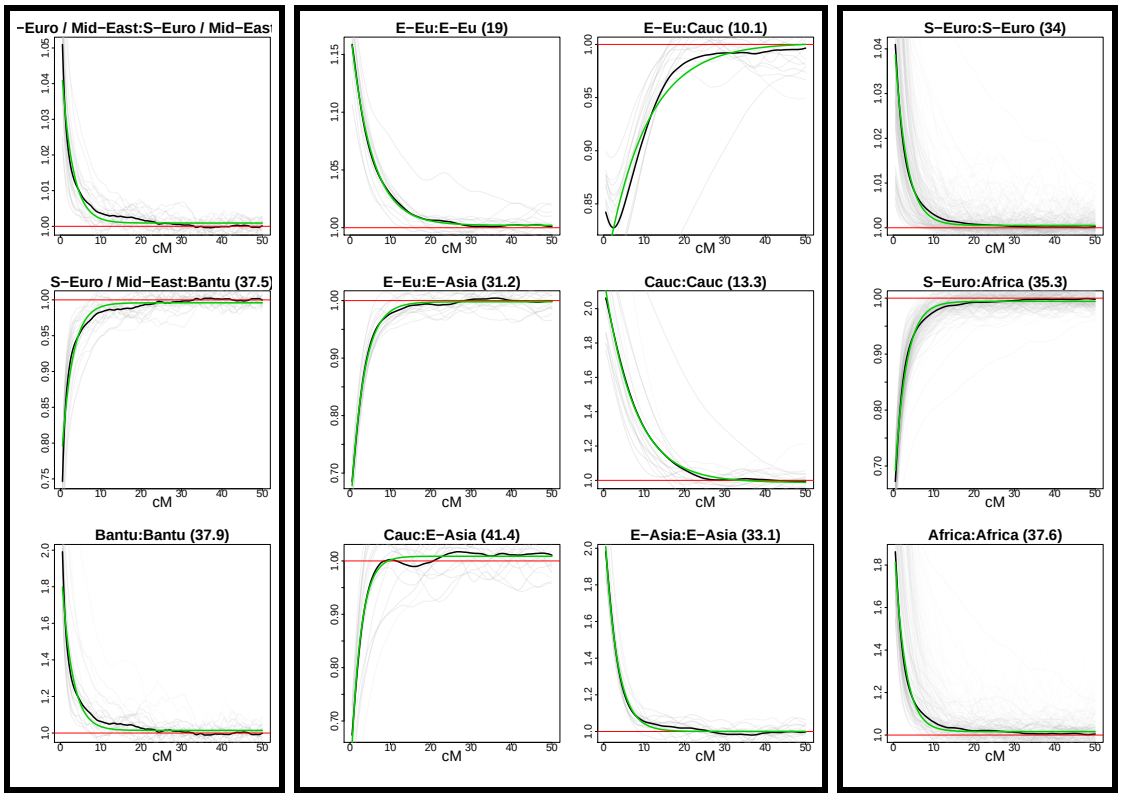


Figure S7: Inferred copying matrices for additional case studies of human admixture based on the HGDP dataset. The copying proportions $\boldsymbol{\mu}_{pa}$ are scaled within columns to % of the most copied donor population so that each cell shading is equal to $100.\boldsymbol{\mu}_{pa}/\arg\max_p \boldsymbol{\mu}_{pa}$. Along the top are the marginal ancestry proportions for each admixed target population.



(a) Moroccan 2-way (b) Chuvash 3-way (c) N-Africa 2-way

Figure S8: Coancestry curves for additional case studies of admixture within the HGDP dataset, corresponding to the copying matrices $\boldsymbol{\mu}$ shown in Figure S7. On the top of each sub-plot approximate geographic descriptions of admixing populations are chosen according to the closest donor panels as measured by $\hat{F}_{\text{st}}$ (see Tables S6 to S7) and the estimated number of generations since admixture between each pair of ancestries is given in brackets. Here SSA stands for Sub-Saharan Africa, S-Eu for Southern Europe, etc.

S6.2 Chuvash 3-way Admixture

Results for a MOSAIC 2-way model applied to Chuvash (an Eastern European population in Siberia) appear in the main text. Here we report results for a 3-way model fit, which has lower but comparable $\mathbb{E}[r^2]$ values (0.54 for 3-way and 0.66 for 2-way). In the 3-way admixture fit, the broadly European-like ancestry of the 2-way analysis has now been separated into European-like and Caucasian-like ancestral groups. Figure S7 depicts the copying matrix and Figure S8b the fitted coancestry curves.

Figure S2a shows the sample density estimate over admixture dates obtained on 500 bootstrap samples of the data (see main paper Section Interpretation of Pairwise Decay Parameters for Multiway Admixture for details on implementation). The pairwise events are (Caucasus + East-Asian), (East-Europe + East-Asian), (East-Europe + Caucasus) in order of mean estimated date from older to more recent; however there is a large overlap in bootstrapped estimated dates for the first two events suggesting that they may be well explained by an event involving an East-Asian like population mixing with both the Caucasus and East-Europe populations, perhaps over a period of time.

Russian	0.0044	Turkish	0.0092	Oroqen	0.032
Belorussian	0.0049	Jordanian	0.011	Yakut	0.033
Polish	0.0049	Iranian	0.011	Mongola	0.036
Lithuanian	0.0064	Armenian	0.011	Xibo	0.038
Hungarian	0.0067	Syrian	0.012	Daur	0.038

Table S7: F_{st} estimates between local ancestries and the closest 5 panels in Chuvash 3-way admixture. The F_{st} estimate between the inferred local ancestries is 1x2=0.021 1x3=0.11 2x3=0.15. The R_{st} is 1x2=0.0049 1x3=0.084 2x3=0.09.

S6.3 North African 2-way: reduced panels

To create the North Africa superset of populations we amalgamated individuals from the following 8 groups: Bedouin (45 individuals), Druze (42), Egyptian (10), Jordanian (18), Moroccan (22), Mozabite (25), Palestinian (46), Tunisian (12) to give a total of 220 individuals. All groups exhibit admixture between Mediterranean and sub-Saharan African-like (14%) ancestral groups when analysed separately (see for example Section S6.1 above). Figure S7 shows the copying rates to the remaining donor groups, with the panels listed above as well as masked from the algorithm. We further restrict the donor panels to only French, English, Scottish, Welsh, Germany-Austria, Ireland, North-Italian, Yoruba, Bantu Kenya, Bantu South Africa, Mandenka, Sandawe, Biaka Pygmy, Hadza, and Mbuti Pygmy. This is to avoid attenuation of the selection signal at the HLA described in the main text due to the potentially widespread occurrence of similar HLA haplotypes across an extended geographic region. Figure S8c shows the

coancesty curves showing a good fit to a single event approximately 35 generations ago. Despite the amalgamation of individuals from different populations, most individually calculated coancestry curves (faint grey lines) are reasonably close to the consensus curve created using all genomes (heavy black line). Table S8 provides F_{st} estimates of the closest matched donor panels to the inferred ancestral mixing groups.

N.Italian	0.0065	BantuKenya	0.019
French	0.0093	Yoruba	0.022
GerAus	0.011	Mandenka	0.023
English	0.013	BantuSA	0.025
Welsh	0.014	Sandawe	0.026

Table S8: F_{st} estimates between local ancestries and the closest 5 panels in North African 2-way admixture. The F_{st} estimate between the inferred local ancestries is 0.11. The R_{st} is 0.16.

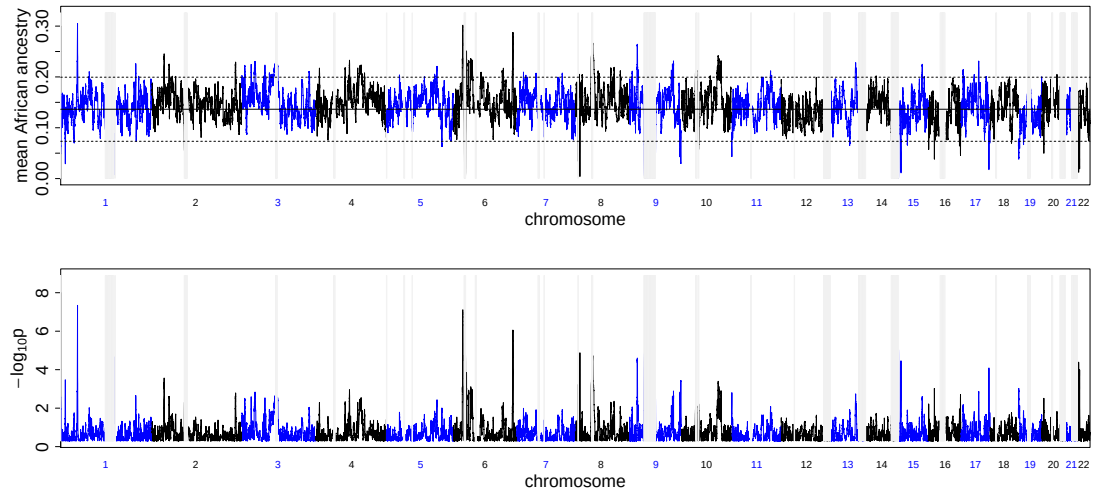
S7 Post-Admixture Selection at the HLA in North Africa?

We explore possible confounding issues that may have created the selection signal at the HLA associated with increased African-like local ancestry amongst North African individuals who are inferred to be admixed European and sub-Saharan African.

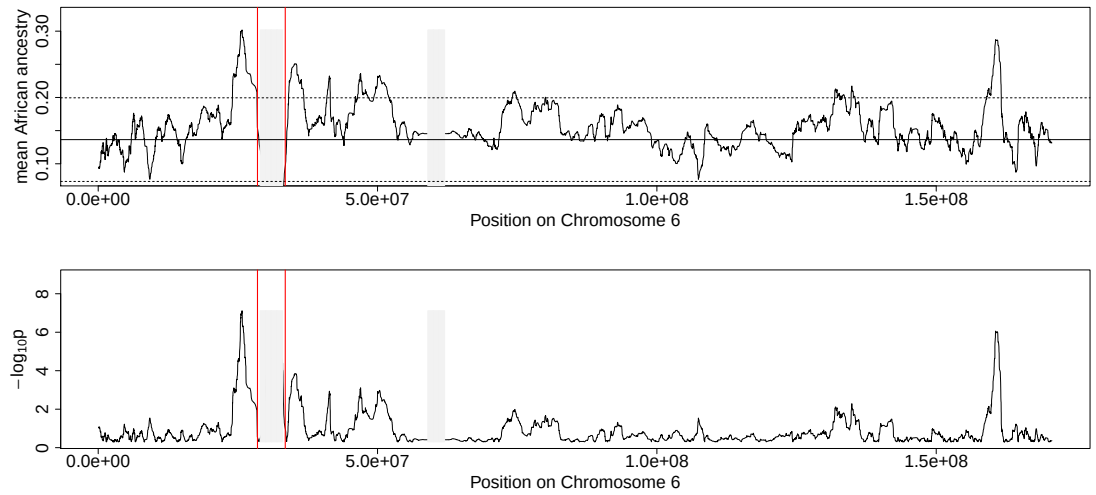
S7.1 Removal of HLA markers

We repeated the selection since admixture in North Africans analysis shown in Figure 8 of the main text, but this time removed all HLA markers. This was done as: (a) The HLA has extremely long-range LD and/or genetic linkage, as well as unusually high diversity due e.g. to balancing selection and high diversity due to balancing selection and (b) if there is a selection effect favouring African ancestral haplotypes post admixture, hitchhiking effects will mean that the regions immediately flanking the HLA should still show a spike in mean African ancestry as estimated by MOSAIC. We therefore expect that any genuine selection signal will be reduced but not entirely removed when all markers from the entire HLA region are removed before running MOSAIC.

Figure S9 shows that this is indeed what we find. There are high and wide spikes in African ancestry flanking the HLA (which is now blanked out as there are no SNPs there in this analysis). These spikes are returning towards the genome wide mean African ancestry within the HLA, because the prior on ancestry moves the inference towards the genome-wide average, in the absence of observed data within the HLA itself.



(a) Mean African Ancestry across all 220 individuals in North Africa against genome position.



(b) Mean African Ancestry across all 220 individuals in North Africa against Chromosome 6 position.

Figure S9: A repeat of the MOSAIC analysis on North African individuals. This time all markers in the HLA region from locus 28510120 to locus 33480577 (on NCBI Build 36.1) were removed before the analysis. This serves as a check for whether the unusual variation patterns (see de Bakker *et al.* (2006)) in the HLA could account for the spike in sub-Saharan African ancestry.

S7.2 Use of all available panels

The full set of donor panels were then used to check if the selection signal was affected. As Figure S10 shows, the selection signal at the HLA is now masked due to preferential copying of HLA alleles extant in Southern European and Middle-Eastern donor populations. Although using the additional donor panels geographically close to Africa yields slightly lower probabilities of African ancestry everywhere, the reduction in the spike in sub-Saharan African inferred ancestry at the HLA loci presents the largest change in across individuals average Africa local ancestry as compared to the analysis using selected donor panels in the main text (a fall of 0.0268 across chromosome 6 outside the HLA versus 0.146 inside the HLA). So HLA alleles are now being copied from close to (but outside) Africa and are inferred as probably non-African ancestry more often.

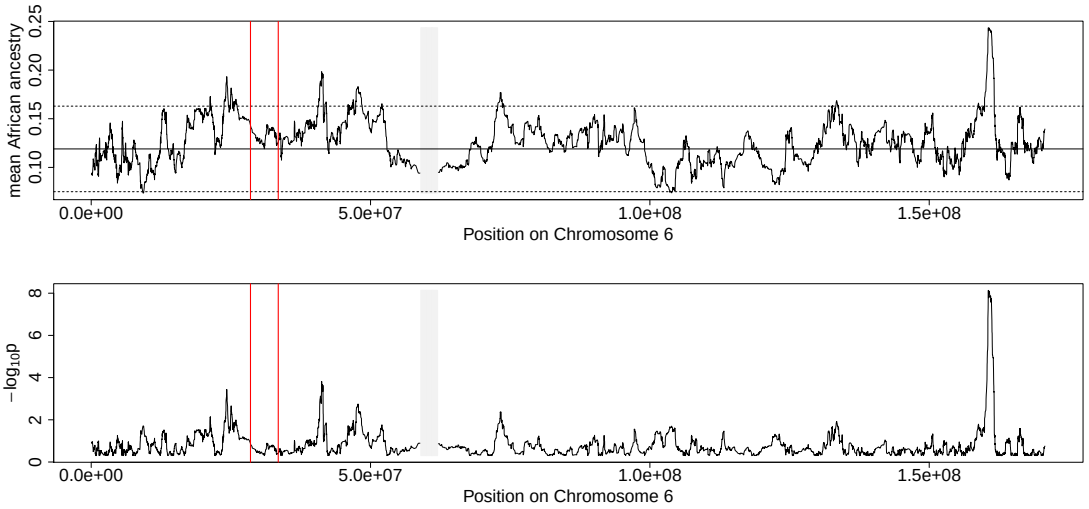


Figure S10: Mean African Ancestry across all 220 individuals in North Africa against Chromosome 6 position when all available global reference panels are used as donors. Figure S11 shows that the apparent elimination of the selection signal at the HLA locus is due to individual target haplotypes copying from populations that are outside Africa but that contain a non-negligible minor African ancestry. This signal reduction could be caused by preferential African-type HLA haplotype sharing across a broad geographic range.

There are two possible explanations for this. The first is that this is "correct" and these are **non-African** ancestry HLA haplotypes from the Middle East and Southern Europe but that they are more similar to African HLAs than northern European ones, so that when MOSAIC doesn't have access to them it copies and infers African HLA alleles. The second explanation is that these are genuine **African** ancestry HLA alleles but MOSAIC will copy them from panels outside Africa (preferentially as they typically flank European tracts and therefore require fewer ancestry switches in the HMM). This is possible if the same sub-Saharan

African set are also over represented in these geographically close-to-African panels due to a similar post-admixture selection effect.

To test for this we then compared within-HLA copying for the reduced and full-panels (all global populations) MOSAIC runs. The selection signal at the HLA now becomes non-significant (p-values range from 0.304 to 0.894), but closer examination of the copying rates suggest support for the second explanation above. We examined the "switching" target North-African HLA haplotypes that are inferred to be African-like in the reduced-panels run but more European-like in the full-panels run. We see from Figure S11 that the panels copied from at the HLA for these switching targets (those that show a 0.5 reduction in the probability of African ancestry, averaged over the HLA) copy mostly from African panels (with French and North Italian also high) but switch to copying a "Mediterranean" set of donors when allowed to do so. This is what changes the inferred ancestry from African-like (minor ancestry) to European-like for these target haplotypes. However, the three most copied from panels are Syrian, South Italian, and Cypriot, all of which are inferred to be admixed between European-like and African-like ancestries (see online browser at `https://maths.ucd.ie/~mst/MOSAIC/HGDP_browser`). Other Mediterranean panels are also copied from at high rates. This suggests that using the expanded reference panel set allows copying of donor panels that are themselves enriched for African (minor) ancestry at the HLA, which might mask a real selection signal.

Note that this is not evidence that MOSAIC cannot be used with imperfect / admixed donor panels (see Section S4 and Section S5 for studies demonstrating robustness to this). It is the presence of a selection signal that is common to highly relevant panels and the target admixed samples that causes the signal to be lost, but this is a local only effect.

S8 Simulation of positive selection

We simulated positive selection in an admixed population for a single locus on Chromosome 6. We begin with ancestry proportions equal to those inferred in North Africa (minor ancestry of 15%, see Section Possible Selection Signal at the HLA in North Africa of main paper). Using $N_e = 10,000$ diploid individuals, we simulated random recombinations along chromosomes of genetic length equal to Chromosome 6 (1.93 Morgans) for 31 generations. The number of recombinations is Poisson with rate 1.93 and the locations of the recombinations are uniform along genetic distance. We assume a Wright-Fisher with selection model of random mating amongst individuals and keep track of the simulated ancestry switch points continuously along each chromosome, as well as the ancestry of each segment. We set a non-zero selection coefficient at a single locus of $s = 0.035$ such that haplotypes containing ancestry of the minor type a at this locus are up-weighted with a relative weight of $1 + s$. i.e. when considering parents, each individual selected parents randomly with the probability of selecting a parent of ancestry

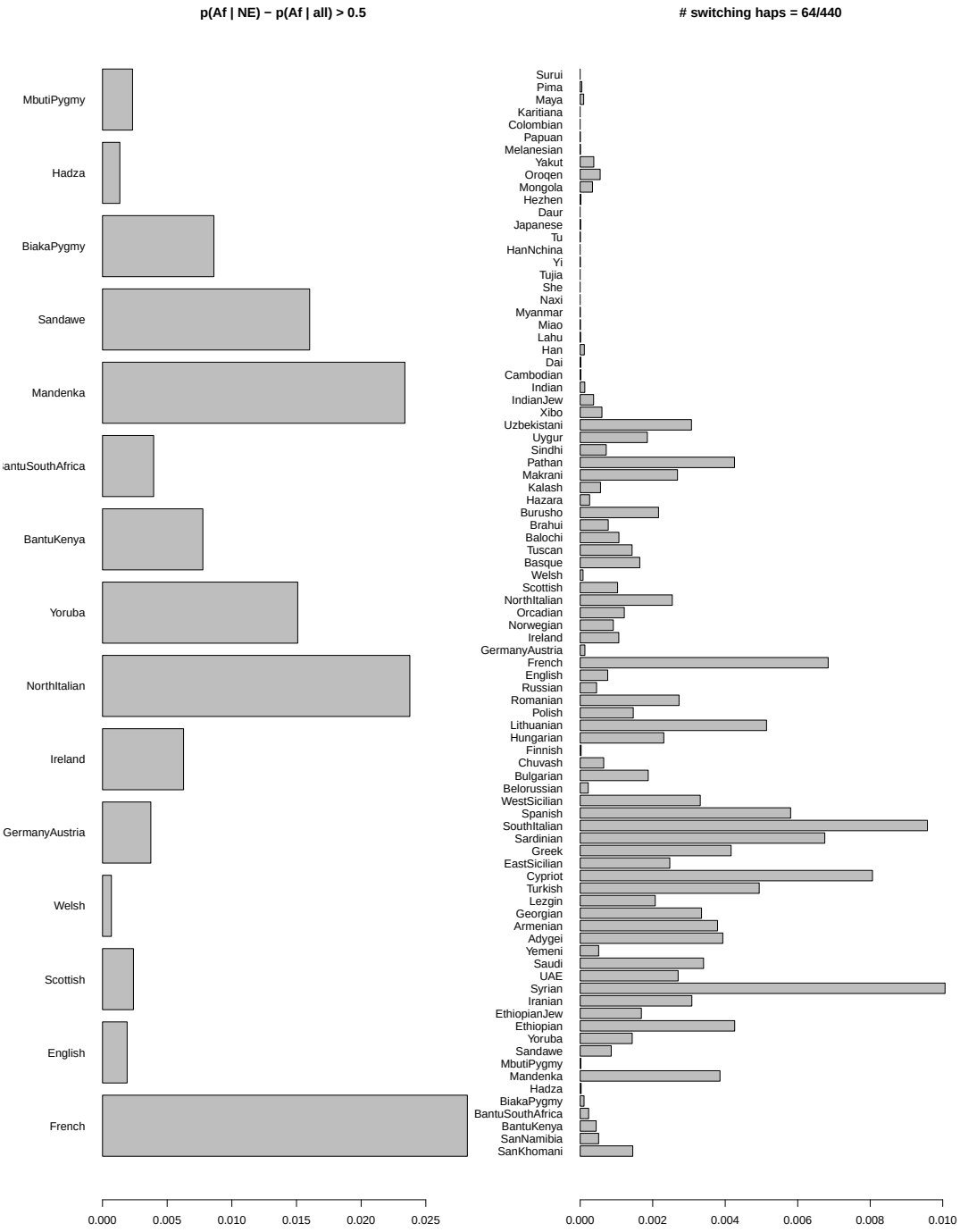


Figure S11: Copying rates for each panel at the HLA averaged across target haplotypes that show a reduction of at least 0.5 in probability of African ancestry between using the reduced set of donor panels and using the full set of donor panels. On the left are the rates of copying across panels for these individuals in the reduced set and on the right is the rates across the full set of panels for these 64 such haplotypes.

a given by $\frac{(1+s)N_a}{2N_e+sN_a}$, where N_a is the number of haplotypes of ancestry a at the selected locus. After 31 generations, we have 10^4 admixed individuals from which we sub-sample 220 diploid individuals (440 haplotypes). We then plot the average ancestry across these individuals against locus in Figure S12, noting the spike centred at the locus simulated to be under selection. We show only true local ancestry here as inferred ancestry is shown to be highly accurate, including at the HLA, in Figure S13 below. The reason for this choice is that MOSAIC simulates only the final generation post admixture by assuming neutral selection of ancestry along the genome.

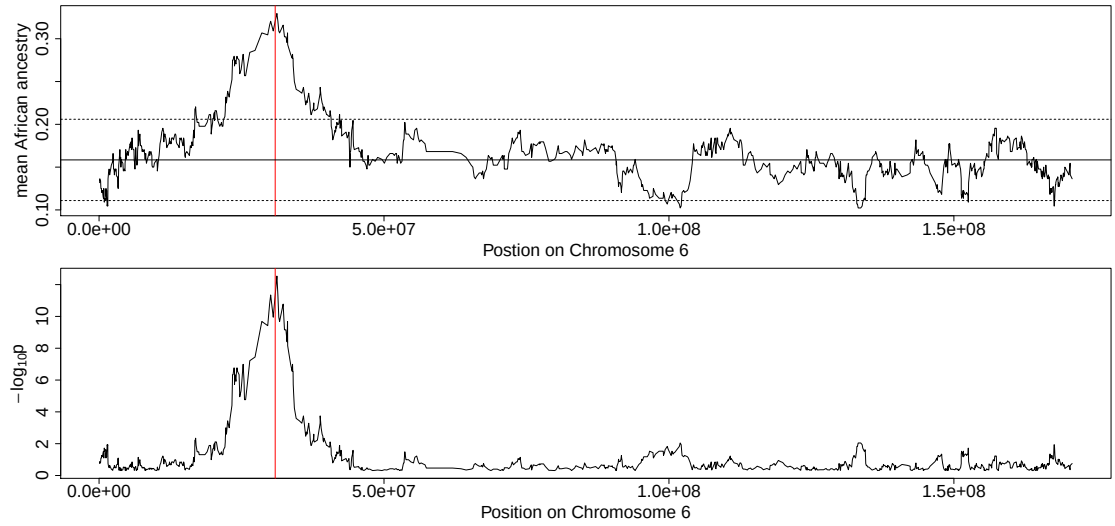


Figure S12: Mean Ancestry in a Wright-Fisher simulation on Chromosome 6 with positive selection at a single locus. There is a highly significant and broad peak at the locus under selection. Note the width of the resulting spike due to hitchhiking of neighbouring loci. The solid vertical line is the mean ancestry outside the selected region and the dashed lines denote ± 2 standard deviations.

Conversely, when we simulate admixture 31 generations ago using 4 real North Italian and Bantu Kenyan diploid genomes (the closest populations to the two ancestral groups as per Table S8) and use the remaining panels (French, English, Scottish, Welsh, Germany-Austria, Ireland, Yoruba, Bantu South Africa, Mandenka, Sandawe, Biaka Pygmy, Hadza, and Mbuti Pygmy) to fit MOSAIC, we do not observe a bias towards African like ancestry at the HLA, as seen in Figure S13. On chromosomes 5 and 6 we simulate 8 admixed haplotypes from the available North Italian and Bantu-Kenyan samples by first sampling ancestry breakpoints from along the genome (ancestry segment lengths are exponentially distributed with a rate equal to 31 Morgans); we then fill in the haplotypes by stitching together North Italian and Bantu Kenyan haplotypes. Each simulated haplotype is assigned two haplotypes from North Italian and two from Bantu Kenyan and we ensure that haplotypes from the same individual are not used in segments sepa-

rated by only one segment from another ancestry as this could bias the method towards ignoring that ancestry switch. In each case, the ancestry proportions are taken such that the expected genome-wide ancestry from the Bantu-Kenyan genomes is 0.14. The simulation is then repeated 20 times, yielding the result in Figure S13. The mean r^2 to true local ancestry across these simulations was 0.96 and within the HLA the r^2 was 0.88.

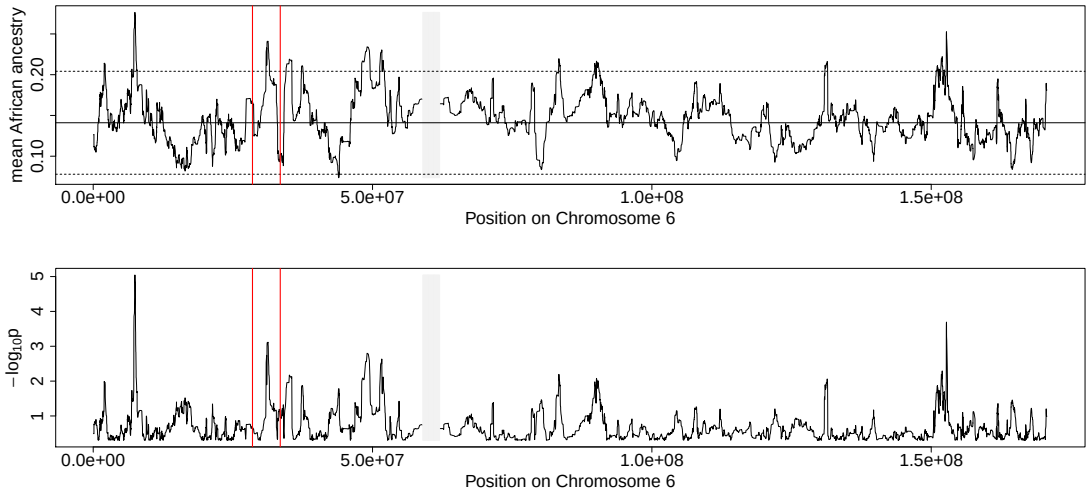


Figure S13: Averaged over individuals local ancestry on chromosome 6 for 20 simulations from MOSAIC. In each run, admixture was simulated using North Italian and Bantu Kenyan genomes 31 generation ago and the other populations listed in Section S6.3 were used as donor panels. Due to the low number of available genomes we cannot recreate the 220 individual scenario of the amalgamated North African study, however no bias towards African-type HLA haplotypes was observed under the repeated simulations with randomly generated admixture in this setting.

References

Baran, Y., B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux, *et al.*, 2012 Fast and accurate inference of local ancestry in Latino populations. Bioinformatics **28**: 1359–1367.

de Bakker, P. I., G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, *et al.*, 2006 A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nature Genetics **38**: 1166–1172.

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

Hellenthal, G., G. B. Busby, G. Band, J. F. Wilson, C. Capelli, *et al.*, 2014 A genetic atlas of human admixture history. Science **343**: 747–751.

Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante, 2013 RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics **93**: 278–288.