Supplementary Texts for the paper:
LINADMIX: Evaluating the effect of ancient admixture events on modern populations

Lily Agranat-Tamir *et al*.

Text S1. Simulating admixed genomes

To test the performance of LINADMIX, we simulated admixed genomes (target populations) in predetermined mixing proportions. We used modern genomes as the basic building blocks of the simulations and down-sampled some genomes to mimic ancient DNA data, as detailed below. We used ADMIXTURE's parameters to prune the SNPs in Plink (--indep-pairwise 50 10 0.1) (Purcell *et al.*, 2007) on the affymetrix Human Origins array data (Lazaridis *et al.*, 2014) in order to avoid dependence between SNPs. After pruning, we ended up with 49,472 autosomal SNPs. We phased the data with the imputation.sanger.ac.uk website using the GRCh37 as reference.

To simulate an admixed genome, we used five individuals from each source population. These individuals were not used to test the performance of the algorithm. For the testing we used different samples from the source populations. For the construction of the admixed genome we selected sequentially chromosomal segments from the sources. Predetermined mixing proportions were used to select the source chromosome of a segment. The length of the segment (in Morgans) was sampled from an exponential distribution with rate equal to the number of generations since admixture. We simulated five independent admixed individuals for each target population.

Down-sampling genomes to model "ancient" DNA data

In typical ancient DNA data, a large proportion of the SNPs are missing, and for non-missing SNPs, only one allele is reported by selecting (at random) a single read covering the SNP (pseudo-haploid data). To simulate genomes from an ancient source, we selected at random SNPs for which genotypic information was considered missing. For the remaining SNPs, we selected one allele at random. In all the models analyzed with LINADMIX, unless otherwise specified, the sources were either simulated "ancient" genomes or true ancient genomes. Table S1.1 lists the number of individuals down-sampled to simulate ancient data in each of the populations used in our simulations.

population	Number of genomes turned into "ancient"
English	4
Jordanian	4
Iranian	3
Russian	4
Spanish	4
German	10
French	25
Palestinian	38
Syrian	8
Somali	13

Table S1.1 The number of individuals in each source population for the LINADMIX model.

Running ADMIXTURE on the simulated genomes

We ran ADMIXTURE on each simulated genome with parameter $K = 6$ (the number of ADMIXTURE ancestral populations). ADMIXTURE's input genomes always included a background of 1,578 genomes of Eurasian and African populations (Patterson et al., 2012, Mallick et al., 2016), to which individuals (turned into "ancient") from the populations used for the simulations were added.

Text S2. Calculation of $\hat{\Sigma}^l$

The empirical Fisher information matrix is calculated from the following joint log-likelihood of the parameters given the data (Alexander et al., 2009):

$$l(Q,F|G) = \sum_{l=1}^{L}\sum_{j=1}^{J}\left\{g_{jl} ln\left[\sum_{k=1}^{K} q^l(k)f_j(k)\right] + \left(2-g_{jl}\right) ln\left[\sum_{k=1}^{K} q^l(k)(1-f_j(k))\right]\right\}. \qquad \text{(i)}$$

Here, L is the number of all individuals in the data for ADMIXTURE, J is the number of SNPs in the data, Q is the matrix of the ADMIXTURE representations (q vectors) of all individuals, F is the ADMIXTURE output for the frequencies of the minor alleles in each ADMIXTURE ancestral population, and $G = \{g_{jl}\}$ is the matrix of genotypes of all individuals l. Elements in G take the values 0, 1, or 2, representing the number of occurrences of the minor allele in SNP j. Because $q^l(K) = 1 - \sum_{k=1}^{K-1} q^l\,(k)$, there are only $K-1$ free parameters for each individual and $q^l(K)$ is replaced by $1 - \sum_{k=1}^{K-1} q^l(k)$ in (i).

The empirical Fisher information matrix is calculated from the second order partial derivatives of (i) with respect to the q parameters. The first order derivative with respect to $q^l(k)$ is

$$\frac{\partial}{\partial q^l(k)} l(Q,F|G) = \sum_j \left\{\frac{g_{jl}\left(f_j(k)-f_j(K)\right)}{\sum_t q^l(t)f_j(t)} + \frac{\left(2-g_{jl}\right)\left(f_j(K)-f_j(k)\right)}{\sum_t q^l(t)\left(1-f_j(t)\right)}\right\}. \qquad \text{(ii)}$$

The Hessian matrix of $l(Q,F|G)$ is obtained by taking partial derivatives of (ii). This matrix is block-diagonal since mixed derivatives that involve different individuals vanish. Denote the block that is associated with individual l by $H^l\left(q^l, F\middle|G\right)$. The (k',k) element of this block is given by:

$$-\sum_j \left(f_j(k')-f_j(K)\right)\left(f_j(k)-f_j(K)\right)\left\{\frac{g_{jl}}{\left(\sum_t q^l(t)f_j(t)\right)^2} + \frac{\left(2-g_{jl}\right)}{\left(\sum_t q^l(t)\left(1-f_j(t)\right)\right)^2}\right\}. \qquad \text{(iii)}$$

Finally, we take $I^l\left(q^l\,, F\right) = -H^l\left(q^l, F\middle|G\right)$.

Text S3. Variance of small samples

The sample variance is an unbiased estimator of the population variance, in the sense that the mean of the sample variance is equal to the true population variance. However, the sampling distribution of the sample variance is positively skewed, and thus a typical draw will be smaller than the mean, underestimating the true variance. This problem is more acute when the variance is estimated from a small sample. Underestimating the population variance of the ADMIXTURE proportions will result in bootstrap simulations that do not fully represent the uncertainty in the data, and can lead to over-optimistic confidence intervals for the mixture coefficients, which is undesired.

We investigated the inaccuracy in the estimation of the variance when the sample size is small. To this end, we considered three populations with relatively large sample sizes: Palestinians (38 individuals), French (25 individuals) and Spanish (48 individuals). For each population, the complete sample variances of the different ADMIXTURE components served as gold standard. This gold standard is marked *yellow* in Figure S3.1. The subject of the investigation was the validity of variance estimators that rely on a sample of size 4.

Two estimation strategies were considered. One strategy used the standard sample variance. This strategy is marked *light blue.* The other strategy, that is marked *dark blue*, added to that variance the inverse of the empirical Fisher information of a random individual. The distributional properties of these strategies were assessed via the empirical distributions of 100 resamples of size 4 each.

The first column of the bar plots in Figure S3.1 presents the expectation of the estimators in both strategies. Comparing the *light blue* and the *dark blue* bars to the *yellow* bars validates that the first strategy is unbiased and the second strategy is positively biased. On the other hand, the second column of bar plots in the figure shows the medians of the same estimators. They suggest that the median in the second strategy is usually closer to the target values than the median in the first strategy.

We conclude that for a small sample the standard sample variance of the components produced by ADMIXTURE is more likely than not to underestimate the true variances. The addition of the inverse Fisher information seems to be a partial remedy.

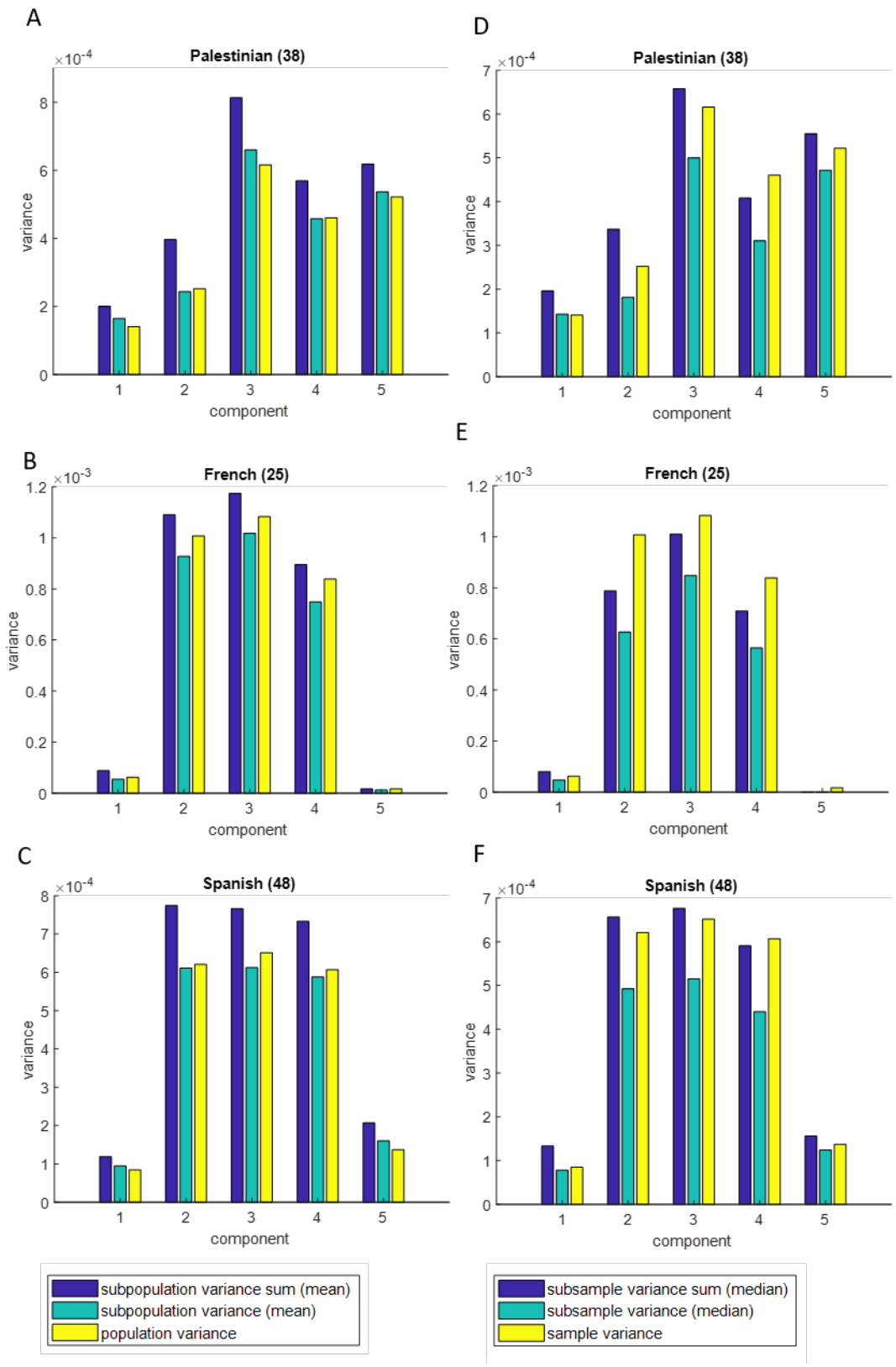

Figure S3.1. Variance of small samples. The means (A-C) and medians (D-F) of the estimated variances of the first five components of the ADMIXTURE representation across 100 random subsamples of size four of three different populations (Palestinian, French and Spanish) using two estimation strategies vs. the estimated variances of the entire sample (size in brackets). Calculation of the subsample variance using only the sample variance is given in blue; Calculation of the subsample variance as the sum of the sample variance with the inverse of the empirical Fisher Information is given in light blue. Calculation of the entire sample variance is given in yellow.

Text S4. Modifications to the estimation of the standard errors

In practice, we include in the bootstrap sampling of an individual only ancestral populations that are sufficiently represented: The variance of ancestral components of q^l that contributed less than 1% was set to zero. Likewise, we set to zero the components and variance of ancestral components of q_i that for all samples from population i had a contribution of less than 1%. In addition, we set to zero negative values of $q_i^{l,(b)}(k)$ and renormalized the vectors to ensure they sum to one.

Text S5. Modification to the calculation of the p-values

Similarly to the computations of the standard errors, we include only ADMIXTURE ancestral populations that contribute sufficiently to the populations in the model. The modifications in the simulation of the source populations are identical to those described in Supplementary Text S4, "Modifications to the estimation of the standard errors". In the simulation of the target population, in addition to the above, we included ADMIXTURE ancestral populations with a sufficient contribution to $q_0^{(b)}$, namely for which $q_0^{(b)} \geq 0.01$. Also, when calculating W and $W^{(b)}$ we only sum over ancestral ADMIXTURE populations k with a sufficient (1% or more) contribution to q_0 (when calculating W) or $q_0^{(b)}$ (when calculating $W^{(b)}$).

Text S6. Segments lengths

We tested the effect of segment length by mixing Jordanians and English in proportions of 0.2:0.8, 0.5:0.5, and 0.8:0.2, as before, but now assuming an admixture event that happened either 20, 50, 80 or 160 generations ago.

All simulations gave similar results, with a maximum error ranging from 4% points (for 20 generations) to 5% points (for 80 generations), and an average error ranging from 2.1% points (for 20 generations) to 2.8% points (for 50 generations) (Figure S6.1, Table S6.1). The average of the standard errors of the mixing coefficients ranged from 5.2% (for 50 generations) to 5.7% (for 80 generations). All simulations consistently produced p-values higher than 0.05, compatible with a model that adequately describes the data (Table S6.1).

We conclude that LINADMIX is not sensitive to the accumulation of recombination events over time, and is therefore robust to changes in time since admixture. This is expected as LINADMIX is based on ADMIXTURE, which in turn considers only allele frequencies at single SNPs and not haplotypes.

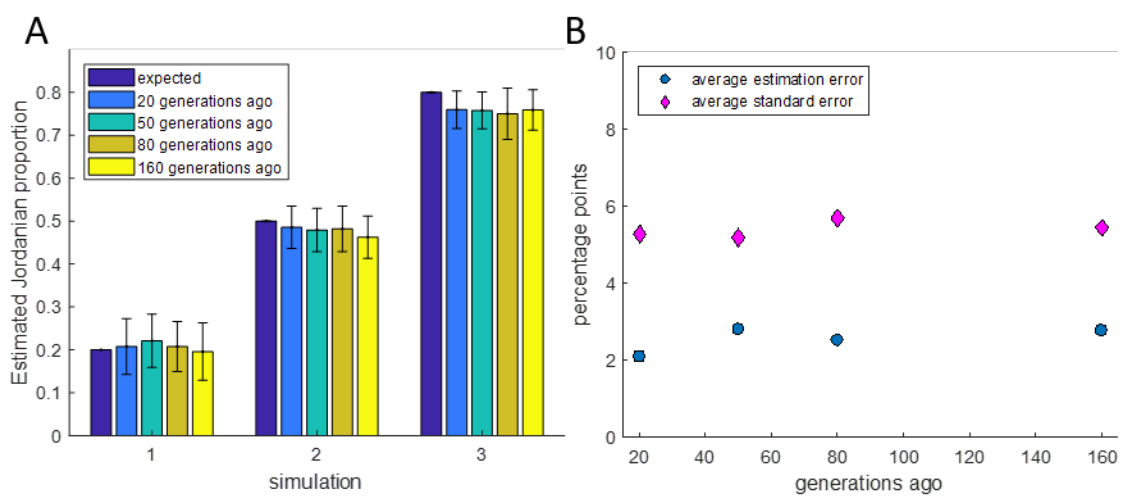

***Figure S6.1**. The effect of the time since admixture on LINADMIX. (A) The estimated Jordanian contribution in simulations where Jordanians and English were mixed in varying proportions, for different times since admixture. (B) LINADMIX performance, measured in several ways, for different times since admixture.*

number of generations ago	average error (percentage points)	maximum error (percentage points)	average standard error (percentage points)	p-value (model 1)	p-value (model 2)	p-value (model 3)
20	2.09	4.05	5.27	0.9307	0.9258	0.6358
50	2.80	4.21	5.17	0.7779	1.0000	0.6017
80	2.52	5.00	5.69	0.4693	0.9883	0.4180
160	2.77	4.13	5.43	0.9548	0.6686	0.6766

Table S6.1*. Simulating different times since admixture.*

Text S7. Finding F_{ST} limits for potential proxies

We tested how genetically remote a population can be and still serve as a proxy to the original source population. To this end, in addition to proxies with F_{ST} values of 0.001, 0.002, and 0.003 relative to the original source, that were analyzed in the main text (section *genetic drift*), we also examined proxies with F_{ST} values of 0.004, 0.005 relative to the original source for the Jordanian-English simulations; and 0.007 for the Jordanian-Iranian simulations. For F_{ST}=0.004 we took Iranians as a proxy to Jordanians; for F_{ST}=0.005 we took Russians as a proxy to English; and for F_{ST}=0.007 we took Iran_ChL as a proxy to Iranians (Supplementary Table S1). We looked at the difference between the average absolute estimation errors of the models with the original source and those for the models with the proxies. In addition, we looked at the p-values of the models (Table 1, Table S4).

For targets that are admixtures of Jordanians and English (F_{ST} = 0.013) we found that up to F_{ST}=0.004 the estimation error is fairly similar between the original and the proxy sources, however the p-values when using the proxy are all below 0.05, leading to a rejection of the model (Figure S7.1). For F_{ST} of 0.005, almost all the p-values are below 0.05, and in addition the average estimation error in the models with the proxies was the highest. Therefore, it appears that populations can serve as proxies (when the target is a mixture of relatively distant populations) when their distance from the original source is no more than F_{ST}=0.003.

For the targets that are an admixture of the closely related Jordanians and Iranians (F_{ST}=0.004) we found that already at F_{ST}=0.003 the estimation error is high compared to that of the original source. Therefore, it is possible to use proxy populations with a distance of up to F_{ST}=0.002 from the original source.

Target: Jordanian-English	
Pairs of populations (original and replacing)	**F_{ST}**
English-French	0.001
English-German	0.001
Jordanian-Syrian	0.002
Jordanian-Palestinian	0.003
Jordanian-Iranian	0.004
English-Russian	0.005
Target: Jordanian-Iranian	
Pairs of populations	**F_{ST}**
Jordanian-Syrian	0.002
Jordanian-Palestinian	0.003
Iranian-Iran_ChL	0.007

Table S7.1 F_{ST} of original source populations and potential proxies. FST values of the pairs of original and potential proxy source populations used to detect the F_{ST} limits.

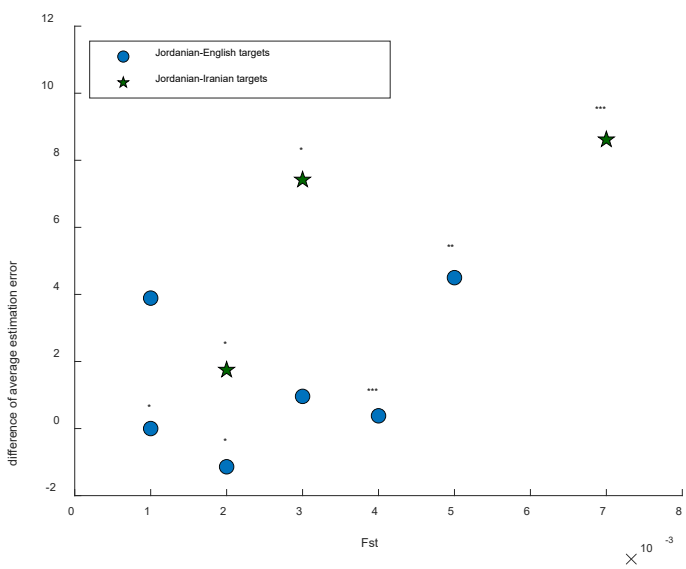

Figure S7.1 Effects of a proxy source population as a function of its F_{ST} compared to the original source. Five pairs of original-proxy sources with four different F_{ST} values (x-axis) were tested for the Jordanian-English simulations; and three pairs of original-proxy sources were tested for the Jordanian-Iranian simulations. The y-axis shows the difference of the average (over the three simulations of Jordanian-English or Jordanian-Iranian) absolute estimation error of LINADMIX between models with the proxy source and models with the original source. For F_{ST}=0.001 the top data point refers to the replacement of English with French. The number of asterisks above points in the graph represents the number of models where the p-value was lower than 0.05 when the proxy was given as a source.

Text S8. Somali-English simulations

To look at target populations that are admixtures of remote populations we mixed Somalis and English (F_{ST}=0.075) at three different mixing proportions, 0.2:0.8, 0.5:0.5, and 0.8:0.2, assuming 50 generations since the admixture event. We looked at models where the sources are in pseudo-haploid form and have 60% missing data. The estimation error was extremely small, less than 1% for all mixing proportions. The standard errors were also low, as expected for targets that are mixtures of genetically distant populations. All p-values were higher than 0.05.

We also tested how proxy populations perform as sources in this scenario, replacing English by either French or Germans. Replacing English with Germans even reduced the estimation errors, however for the 80% English target population the p-value was slightly below 0.05 (0.0406). Replacing the English with French slightly increased the estimation errors, in less than 1.5%. Therefore, as in the other simulations, LINADMIX is robust to drift.

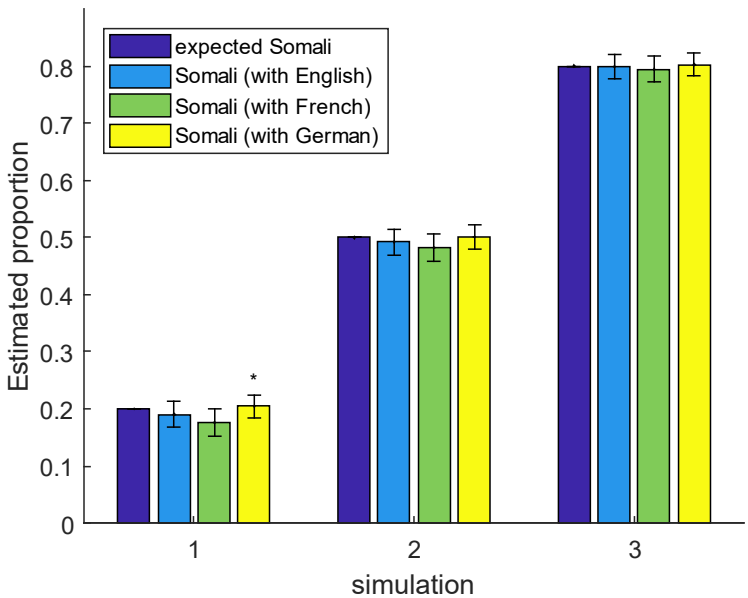

Figure S8.1. Admixture of remote populations. Models include the original sources (English and Somali), as well as proxies to English (either German or French). The asterisk represents that the p-value of the model was below 0.05.

target 1: Somali 20% English 80%
target 2: Somali 50% English 50%
target 3: Somali 80% English 20%

sources provided to LINADMIX	average error (percentage points)	maximum error (percentage points)	average standard error (percentage points)	p-value (target 1)	p-value (target 2)	p-value (target 3)
Somali and English	0.62	0.97	2.20	0.2190	0.3332	0.0991
Somali and French	1.57	2.37	2.40	0.2232	0.3269	0.1156
Somali and German	0.25	0.43	2.09	0.0406	0.2768	0.1639

Table S8.1. Admixture of remote populations. Models include the original sources (English and Somali), as well as proxies to English (German and French). P-values under 0.05 are shaded.

Text S9. The effect of ADMIXTURE perturbations on model fit

In Agranat-Tamir et al. (2020) we looked at the effect of ADMIXTURE perturbations on the estimation of LINADMIX and of model fit in terms of (squared) norms of residuals, that is $\|Q_s\hat{\alpha} - q_t\|_2^2$ where Q_s is the matrix of the admixture representations of the source populations in the model; $\hat{\alpha}$ is the vector of the estimated mixing coefficients; and q_t is the admixture representation of the target population in the model. Here we complete the analysis by looking at the effect of ADMIXTURE perturbations on model fit in terms of the p-value.

The analysis is of 17 modern populations modeled as mixtures of 4 ancient populations. The original analysis is of 1663 individuals with K=6 (this K is the optimal one from the cross-validation procedure suggested in the ADMIXTURE manual). We performed two perturbations: looking at a different not optimal K – K=7; and adding apparently non relevant populations to the set given to ADMIXTUERE, ending with 3515 individuals and a new optimal K of 22. To summarize we found that changing the K from an optimal one to a sub-optimal affects the estimation of LINADMIX more than the addition of individuals: The average absolute difference in the LINADMIX estimation across the 17 target populations and 4 source populations is 6.1 percentage points when K is only sub-optimal and 3.7 percentage points when the set of individuals is increased. The standard errors were very similar in all analyses (average difference of less than 2 percentage points).

When looking at model fit we found in Agranat-Tamir et al. (2020) that the norms were much higher when K was sub-optimal. The addition of individuals changed to some extent the pattern of norms across the 17 models. We now add p-value calculations (Table S9.1). For the original analysis of 1663 individuals and K=6 three were below the 0.05 threshold. (All non-Southern Levant target populations.) As expected from the norms the p values when K is not optimal (K=7) were lower with 11 out of the 17 below 0.05 (intersecting two of the three when K=6). The p-values for the larger set of input individuals to ADMIXTURE were also lower than those with the smaller set – 9 out the 17 were below 0.05 intersecting with all of those in the original analysis. It is important to note that the addition of populations to the ADMIXTURE input resulted in a new, much higher, optimal K. Therefore, it is expected that it would be harder to get model fit.

In conclusion, LINADMIX estimation is quite robust to changes in the ADMIXTURE run, though more so to changes in the input set than to a non-optimal K. The main differences are in model fit tests. Using a non-optimal K is not advisable. In addition, using a large set of individuals that results in a high optimal K is also less advisable as it makes it harder to get model fit. We suggest using a set of individuals that would cover those important for the analysis but that would not have a high optimal K.

	p-value		
target populations	**1663 individuals K=6**	**1663 individuals K=7**	**3515 individuals K=22**
Ashkenazi_Jew	0.1749	0.0001	0.0455
Bedouin_A	0.9717	0.0361	0.7490
Bedouin_B	0.4851	0.3668	0.0001
Druze	1	0.0001	0.0005
Egyptian	0.9036	0.0004	0.3872
English	0.0287	0.0858	0.0197
Iranian	0.0002	0.0001	0.0487
Iranian_Jew	0.5906	0.0001	0.0089
Ethiopian_Jew	0.9999	0.8335	0.7754
Jordanian	0.5536	0.0187	0.1930
Lebanese	0.6138	0.1117	0.2286
Moroccan	0.1827	0.0121	0.0124
Moroccan_Jew	0.1811	0.0001	0.0138
Palestinian	0.9835	0.0013	0.3532
Saudi	0.7233	0.6241	0.4927
Syrian	0.8576	0.1417	0.5590
Tuscan	0.0349	0.0001	0.0002

Table S9.1 Model fit with ADMIXTURE perturbations. P-values below 0.05 are shaded.