An explanation for the sister repulsion phenomenon in Patterson's f-statistics

Gözde Atağ[1], Shamam Waldman[2], Shai Carmi[3] and Mehmet Somel[1]

[1] Department of Biological Sciences, Middle East Technical University, 06800 Ankara, Turkey
[2] Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.
[3] Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem 9112102, Israel

Abstract

Patterson's f-statistics are among the most heavily utilized tools for analysing genome-wide allele frequency data for demographic inference. Beyond studying admixture, f_3 and f_4 statistics are also used for clustering populations to identify groups with similar histories. However, previous studies have noted an unexpected behaviour of f-statistics: multiple populations from a certain region systematically show higher genetic affinity to a more distant population than to their neighbours, a pattern that is mismatched with alternative measures of genetic similarity. We call this counter-intuitive pattern "sister repulsion". We first present a novel instance of sister repulsion, where genomes from Bronze Age East Anatolian sites show higher affinity towards Bronze Age Greece rather than each other. This is observed both using f_3- and f_4-statistics, contrasts with archaeological/historical expectation, and also contradicts genetic affinity patterns captured using PCA or MDS on genetic distances. We then propose a simple demographic model to explain this pattern, where sister populations receive gene flow from a genetically distant source. We calculate f_3- and f_4-statistics using simulated genetic data with varying population genetic parameters, confirming that low-level gene flow from an external source into populations from one region can create sister repulsion in f-statistics. Unidirectional gene flow between the studied regions (without an external source) can likewise create repulsion. Meanwhile, similar to our empirical observations, MDS analyses of genetic distances still cluster sister populations together. Overall, our results highlight the impact of low-level admixture events when inferring demographic history using f-statistics.

Introduction

Over the last decade, Patterson's f-statistics (Patterson *et al.* 2012) have become one of the most commonly applied tools in population genetics studies involving the analysis of genome-wide data. Various f-statistics (f_2, outgroup-f_3, admixture-f_3, and f_4) quantify divergence, shared genetic drift and treeness among populations by comparing allele frequencies. These measurements enable the identification of population structure, differentiation and admixture events and help decipher the demographic history of the studied populations (Peter 2016; Harris and DeGiorgio 2017; Lipson 2020; Peter 2022). The description of f-statistics as being robust to heterogeneity of sample size/coverage and ascertainment biases (Patterson *et al.* 2012) has gained them high popularity.

The outgroup-f_3 statistic in the form of f_3*(Outgroup; Pop2, Pop3)* measures shared drift between *Pop2* and *Pop3* relative to an outgroup, and its magnitude has been used as a measure of pairwise population affinity. The f_4-statistic measures treeness, such that f_4*(Outgroup, Pop1; Pop2, Pop3)* is expected to be 0 if *Pop1* shares the same amount of drift with *Pop2* and *Pop3*, i.e. *Pop1* is a real outgroup to the latter two populations. Significant deviation from 0 can be interpreted as evidence for admixture (Patterson *et al.* 2012), but the f_4-statistic can also be used for studying population structure and shared demographic histories. If *Pop2* and *Pop3* share more of their demographic history than they do with *Pop1*, i.e. shared drift and/or similar admixture events that *Pop1* did not experience, and if there is no post-split admixture, we expect f_4*(Outgroup, Pop1; Pop2, Pop3)* to be non-significant, while f_4*(Outgroup, Pop2; Pop1, Pop3)* and f_4*(Outgroup, Pop3; Pop1, Pop2)* to both be positive. Gene flow from *Pop1* into *Pop3* in the same model will also cause f_4*(Outgroup, Pop1; Pop2, Pop3)* to be positive. Thus, in addition to testing admixture, reciprocal affinity in f_4-tests has frequently been used as a tool for clustering populations based on allele frequency similarities, where clustering implies similar histories. The outgroup-f_3 statistic has similarly been used to evaluate shared history between populations.

Nevertheless, f-statistics are not always straightforward to interpret. Complications of diverse types and origins have been noted, such as possible misinterpretation of an outgroup as an admixture source when using admixture-f_3 (Patterson *et al.* 2012), or the sensitivity of f_4-statistics to SNP ascertainment schemes as showcased for demographic modelling of African human populations using SNPs identified in non-African groups (Bergström *et al.* 2020; Flegontov *et al.* 2023). The most widespread type of issue arguably involves the interpretation of f_4 -statistics under the assumption of a simple demographic model where the real history is substantially more complicated. For instance, unrecognised genetic structure can be misinterpreted as admixture events

(Durand *et al.* 2011; Eriksson and Manica 2012). A similar confounding effect can be created by backward gene flow, as shown for Neanderthal admixture estimates varying spuriously over time due to African gene flow into Eurasia (Petr *et al.* 2019a).

Previous work using f-statistics has noted another unexpected pattern that has yet remained largely unexplored, to our knowledge. Using f_4-statistics, Kılınç et al. observed higher affinity of all Anatolian Neolithic groups to a Neolithic genome from Greece than to each other, a pattern that appeared difficult to interpret (Kılınç *et al.* 2017). Even more surprising was the observation that modern-day North African populations showed higher affinity to modern-day Sardinians than to each other using f_3- and f_4-statistics (Rodríguez-Varela *et al.* 2017). This pattern was hypothesized to stem from gene flow from sub-Saharan Africa into North Africa, but the case was not studied further.

Here, we provide another case of this pattern, which we call "sister repulsion", by comparing genomes from the Bronze Age (BA) East Mediterranean. We then describe a simple theoretical model that can explain this effect, propose different scenarios involving external and regional gene flow to explain sister repulsion, and recapitulate the observed f-statistics using simulations under these models, with various conditions. Our results showcase how complex demographic histories can lead to counter-intuitive patterns of f_3- and f_4-statistics when used for studying shared population history.

Results and Discussion

Our initial motivation was to investigate population structure and admixture patterns in the East Mediterranean during the Bronze Ages, a period which saw population growth and extensive interregional trade networks being established, as inferred from material culture data (Şahoğlu 2005; Kouka 2016; Massa 2016). We studied genetic affinities between published genomic data from n=6 BA populations from modern-day Greece (2400-1070 BCE) and n=5 populations from BA East Anatolia (3700-1303 BCE) (Lazaridis *et al.* 2017; Skourtanioti *et al.* 2020; Lazaridis *et al.* 2022b) (Table S1, Figure S2). We restricted the analysis to 1240K capture data to avoid confounding by technical factors (Margaryan *et al*. 2020) likely caused by reference bias (Rohland *et al.* 2022; Koptekin *et al.* 2023). We note that the results described henceforth are also observed using Central and West Anatolian BA populations, which we do not show here; we instead focus on East Anatolia as analyses involving populations from this region are the most counter-intuitive.

First, we explored genetic clustering among populations with Principal Components Analysis (PCA) and Multidimensional Scaling (MDS) analysis. In the PCA, where the ancient genomes are projected on a PC space calculated from modern-day West Eurasians, genomes from modern-day Greece cluster separately from those from East Anatolia (Figure S1). Then, we calculated the genetic distance between all pairs of populations using 1 - f_3*(Yoruba; Pop1, Pop2)* to represent the distance between populations *Pop1* and *Pop2*. Applying MDS analysis on the resulting pairwise distance matrix (Fu *et al.* 2016), the populations showed a tendency to cluster by region in MDS space, i.e. Greece and Anatolia separated along Coordinate 1 (Figure 1A,B).

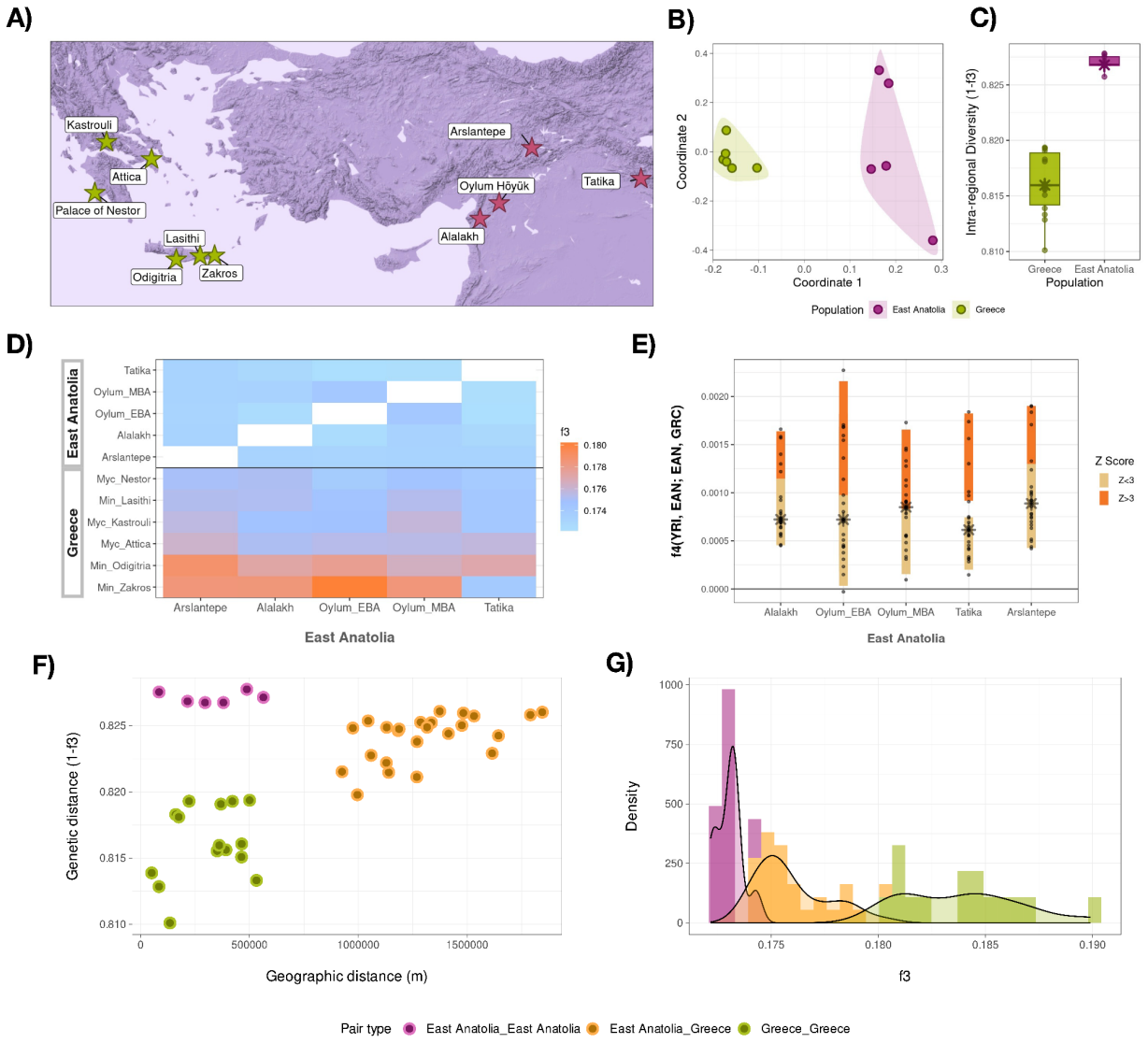


Figure 1: (A) Geographical distribution of the studied Bronze Age genomes from modern-day Greece and East Anatolia. **(B)** Genetic clustering of the populations in line with geography revealed by multidimensional scaling analysis (MDS), using 1 - outgroup-f_3 values as genetic distances. Figure S1 also presents a PCA, which also shows the absence of visible structure in these gene pools. **(C)** Intra-regional diversity of the populations measured as pairwise 1 - outgroup-f_3 values within regions. Each point represents comparisons of settlements from the same region. **(D)** f_3 statistics of the form f_3*(Yoruba; East Anatolia, X)* where X corresponds to populations from modern-day Greece and East Anatolia. **(E)** f_4 statistics of the form f_4*(Yoruba, East Anatolia; East Anatolia, Greece)*, where positive values depict higher affinity of BA East Anatolian

populations towards contemporary populations from modern-day Greece. **(F)** Geographic versus genetic distances (1-f_3) between and within regions. **(G)** Distribution of f_3 values between and within regions.

Intriguingly, however, f_3 statistics revealed a counter-intuitive trend. In contrast to the patterns observed in PCA and MDS, the values f_3*(Yoruba; EAnatoliaY, GreeceX)* were consistently higher than the values f_3*(Yoruba; EAnatoliaY, EAnatoliaZ)*, where *GreeceX* is one of the six populations from modern-day Greece, while *EAnatoliaY* and *EAnatoliaZ* are any of the five populations from East Anatolia (Figure 1B-E,G, Table S2). At face value, this would imply more similar histories between any East Anatolian population with those from modern-day Greece than other neighbouring East Anatolian populations. Plotting the genetic versus geographic distances, we observed a prominent lack of relationship within the East Anatolian populations (Figure 1F).

We next calculated f_4-statistics of the form f_4*(Yoruba, AnatoliaY; AnatoliaZ, GreeceX)*. In 99% of 120 comparisons we observed positive values (42% significant at Z>3, without correcting for multiple testing) (Figure 1D, Table S3). No f_4-statistic was significantly negative, i.e. no East Anatolian group significantly shared more drift with one of its neighbours over a population from modern-day Greece. This starkly contrasts with the expected geographical and genetic clustering [given multiple observations of spatial population structure in ancient humans (e.g. Altınışık, Kazancı, *et al.* 2022; Antonio *et al.* 2024)], while it does mirror the outgroup-f_3 results.

We then asked whether population or time structure within the two regional groups could be behind this pattern, as the sampled individuals span several thousands of years. We investigated this using f_4-statistics within one site in East Anatolia, Alalakh. We randomly split the n=25 genomes from Alalakh into two groups and calculated f_4*(Yoruba, Group1; Group2, Greece)*, ten times independently. We observed positive f_4 values, or sister repulsion, in all ten Alalakh replicates (all f_4 values were significant except for one trial) (Table S4).

Assuming this is not caused by technical artifacts, such as contamination, there could be two explanations for this sister repulsion pattern. One may be gene flow from modern-day Greece into different East Anatolian populations or the ancestors of all East Anatolian populations. An alternative explanation and a possibly more likely scenario could involve admixture from an external and diverse source impacting East Anatolia. This would be similar to that hypothesized by Rodríguez-Varela and colleagues (2017) and analogous to the “outgroup case” described by Patterson and colleagues (2012).

The ancestral components of both Anatolia and Greek Early/Middle Bronze Age genomes comprise Neolithic Anatolia and eastern-related ancestry from Caucasus/Mesopotamia (Clemente *et al.* 2021; Lazaridis *et al.* 2022b; Skourtanioti *et al.* 2023). We hypothesized that additional external gene flow into East Anatolia (e.g. from the Levant, Zagros, or Mesopotamia) that BA populations from Greece did not receive could lead to differentiation within the former region and might explain the observed sister repulsion pattern. As a first test of this idea, we calculated intra-regional diversity within each site, using the pairwise statistic 1 - outgroup-f_3 among genomes within a site as the diversity measure. This revealed significantly higher median diversity in BA East Anatolia than in BA Greece (Wilcoxon rank sum test p=3e-09, Figure 1C, S3).

This supports the hypothesis that gene flow from a divergent source (so-called "deep ancestry") might explain this pattern. We investigated the idea further using the scenario described in Figure 2. We define A and B as two sister populations, X as a more distant lineage, Z as the source of external gene flow into A and B, and O as outgroup to all. Here, A, B and X represent our focal populations; A and B could represent two BA East Anatolian groups, X could be a BA population from the region of modern-day Greece, and Z would be an unknown, distant source that admixes only with East Anatolia. The ancestors of multiple populations are denoted by their merged names (i.e. the ancestor population of A and B is called AB). Note that in the model, A and B are sisters; their demographic histories are the same except for receiving independent gene flow from Z.

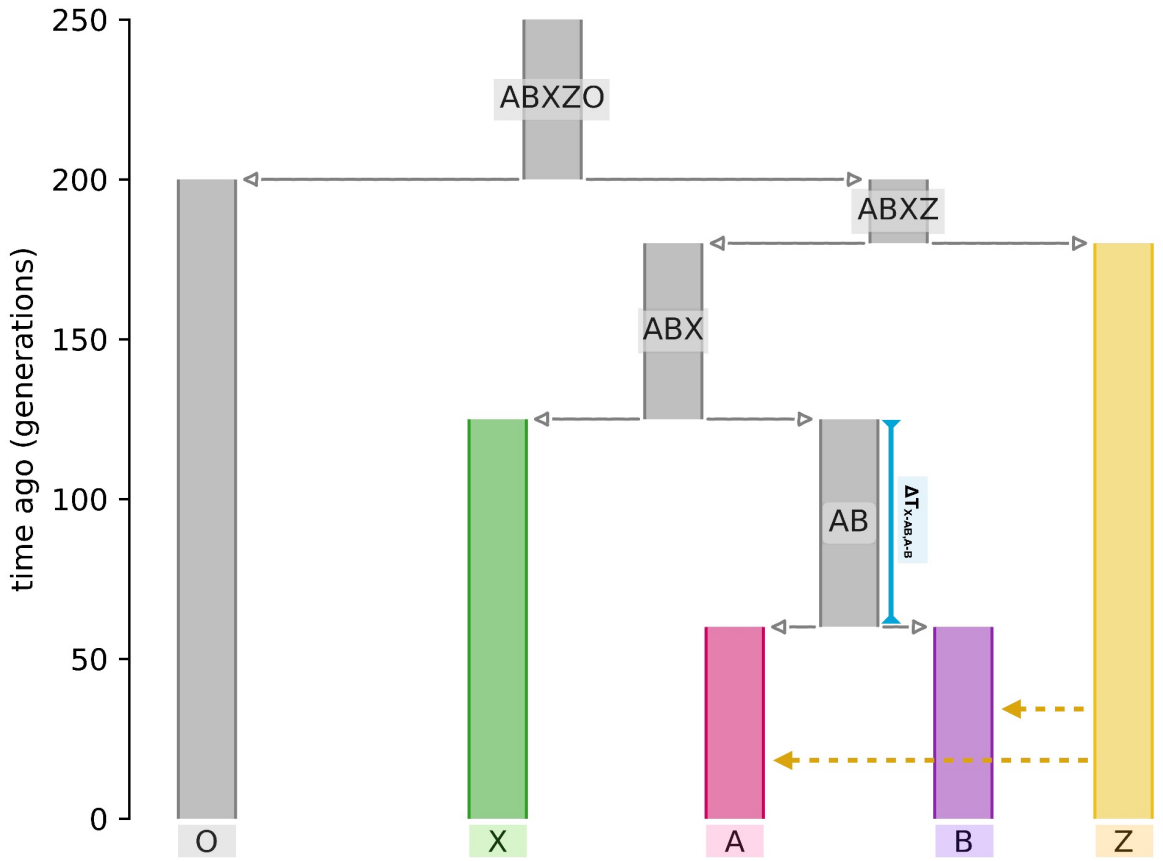
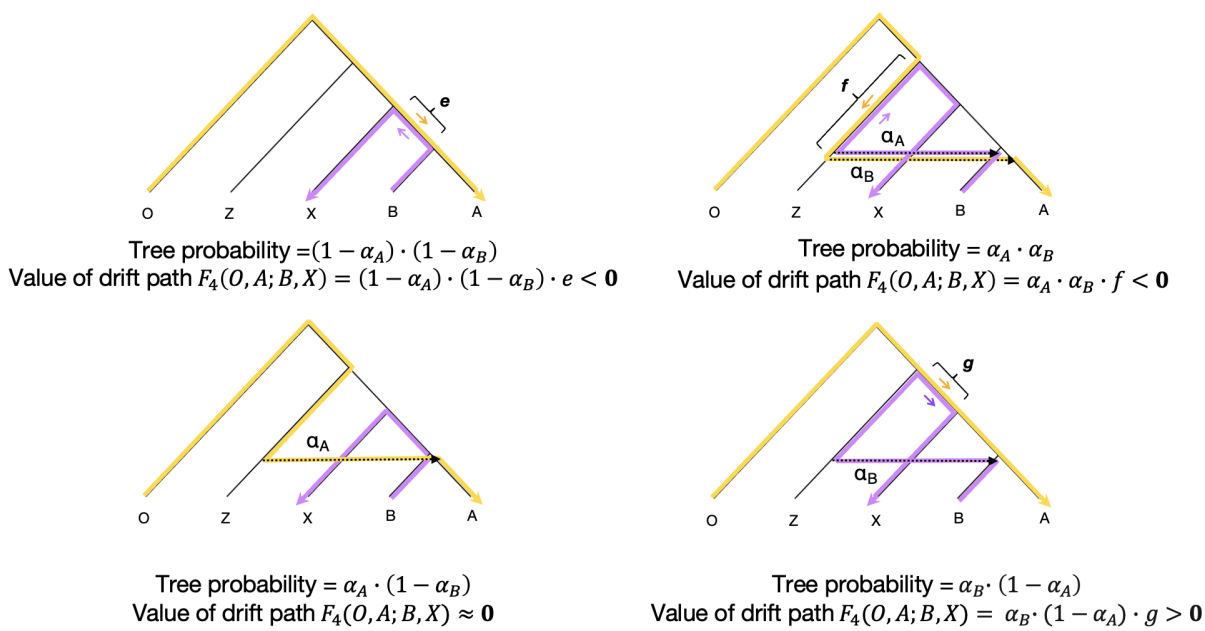


Figure 2: Demographic model used for the population genetic simulations. The ancestral populations are named as combinations of the descending population names. Population O is the outgroup used in the calculation of f-statistics. Dashed lines represent equal rates of continuous migration from Z to A and B per generation after the split between A and B and the present day. The blue line corresponds to the parameter $\Delta T_{X\text{-}AB,A\text{-}B}$.

We studied the theoretical conditions for the expected value of $F_4(O, A; B, X)$ to be positive in this model (we follow Patterson *et al.* in using upper case letters for expected values). The F_4 statistic can be decomposed into the sum of overlapping dirft paths between O to A and B to X. Under gene flow from Z into A and/or B, there emerge four alternative drift paths, described in Figure 3. These depend on α_A and α_B, the probabilities of gene flow (admixture proportions) from Z into A and B. The expected F_4 is the weighted sum of the values of these paths. One of these paths produces a positive value, i.e. sister repulsion, which depends on the branch length g separating Z from the common ancestor of A, B, and X (branch ABX in Figure 2), on the branch length e separating X

from the common ancestor of A and B (branch AB in Figure 2) and also on the admixture proportions α_A and α_B. Note that the branch lengths represent drift and will be a function of time and N_e.

Now let us we assume that the admixture proportions in A and B are the same ($\alpha_A = \alpha_B$) and small (so that α^2 is negligible). Then, $g \cdot \alpha > e$ would fulfill the condition of a positive F_4, i.e. sister repulsion. For instance, in a simplistic scenario where all populations had the same N_e and the admixture proportions α were the same, if BA populations from the regions of modern-day Greece and E Anatolia had split 30 generations ago (~900 years), the latter received gene flow from a population that had split 2000 generations ago (~60,000 years), an admixture proportion of >0.015 would suffice to create a positive F_4 value.



Total value of drift path:

$$F_4(O,A;B,X) = [\boldsymbol{\alpha_B} \cdot (\mathbf{1}-\boldsymbol{\alpha_A}) \cdot \boldsymbol{g}] - [(1-\alpha_A)\cdot(1-\alpha_B)\cdot e] - [\alpha_A \cdot \alpha_B \cdot f]$$

Figure 3: Expected F_4 values under distant admixture. The four trees with arrows show four possible drift paths that can contribute to F_4(O, A; B, X). Here, sister populations A or B have independently received gene flow (dashed arrow) from the distant branch Z with admixture proportions α_A and α_B, respectively, after their split from each other. The probability of each path and its value as a product of its probability and the drift magnitude (indicated by letters *e*, *f*, and *g*) are shown below each tree. The expected average F_4 value (Patterson et al. 2012) is shown at the bottom, with the positive

component in bold. The small arrows indicate the direction of the drift paths from O->A and B->X, which determine the sign of that F_4 path.

Having shown sister repulsion is possible theoretically, we next studied the behaviour of f-statistics numerically with population genetic simulations. Using the general model in Figure 2 we ran multiple simulations where we varied the parameters $m_{Z\text{->}A.B}$, Ne_Z, Ne_{AB}, Ne_{ABX}, $T_{A\text{-}B}$, and $T_{X\text{-}AB}$ (Table 1). The continuous migration $m_{Z\text{->}A.B}$ was set to start subsequent to the A-B split. For simplicity, we refer to $T_{A\text{-}B}$ and $T_{X\text{-}AB}$ as a single variable that shows $T_{X\text{-}AB}$ - $T_{A\text{-}B}$, which we call $\Delta T_{X\text{-}AB,A\text{-}B}$. In each case we sampled 25 genomes each from O, X, A, and B, and calculated various outgroup-f_3 and f_4-statistics.

Table 1: Parameter values used for the main population genetic simulations. Migration rates are given per generation, divergence times are given in generations.

Parameter	**Values**	**Definition**
$m_{Z\text{->}A.B}$	0, 0.001, 0.005, 0.01, 0.1	Migration rate from Z to A and B
Ne_Z	5k, 15k, 35k	Effective population size of source Z
Ne_{AB}	5k, 15k, 35k	Effective population size of ancestral AB
Ne_{ABX}	1k, 3k, 5k	Effective population size of ancestral ABX
$T_{A\text{-}B}$	30, 60, 90	Divergence time of A and B
$T_{X\text{-}AB}$	100, 125, 150	Divergence time of X and AB
$\Delta T_{X\text{-}AB,A\text{-}B}$	10, 65, 120	Difference between $T_{A\text{-}B}$ and $T_{X\text{-}AB}$
Ne_X*	1k, 3k, 5k	Effective population size of X
$m_{X\text{->}A,B}$*	0, 0.001, 0.005, 0.01, 0.1	Migration rate from X to A and B

* The given parameter values for Ne_X and $m_{X\text{->}A,B}$ were tested only for scenarios where Ne_{ABX} = 1k and $\Delta T_{X\text{-}AB,A\text{-}B}$ = 65. In the remaining scenarios, Ne_X was fixed at 5k and $m_{X\text{->}A,B}$ at 0.

In this model (Figure 2) one would naively anticipate $f_4(O, A; B, X)$ values to be negative since the populations A and B appear to share more similar histories with each other than with X, as in tree A in Figure 3. Indeed, in the absence of gene flow from Z into A/B

(m=0), f_4 statistics are mostly negative, reflecting higher shared drift between A and B than with X (Figure 4, Figure S2). However, f_4 values change as we allow migration from Z into A/B (from m=0 to m=0.1), although in non-linear fashion. When we increased migration from 0 to 0.01, we observed a monotonic rise in $f_4(O, A; B, X)$ values, i.e. an increasing trend towards sister repulsion (Figure 4). Under specific conditions, e.g. Ne_{ABX} = 1k, $0.001 \leq m < 0.1$ and $\Delta T_{X\text{-}AB,A\text{-}B} = 10$, all $f_4(O, A; B, X)$ values are positive (Figure 4, S4). The same applies to $f_4(O, B; A, X)$. This sister repulsion pattern is analogous to that observed in the empirical data where two East Anatolian populations showed higher affinity to populations from modern-day Greece than to each other (Figure 1D). We also repeated our random split approach explained above for the real data in a simulated scenario for individuals in population A and again obtained all positive f_4 values (Table S5). In summary, we can replicate the empirical result effectively under specific conditions.

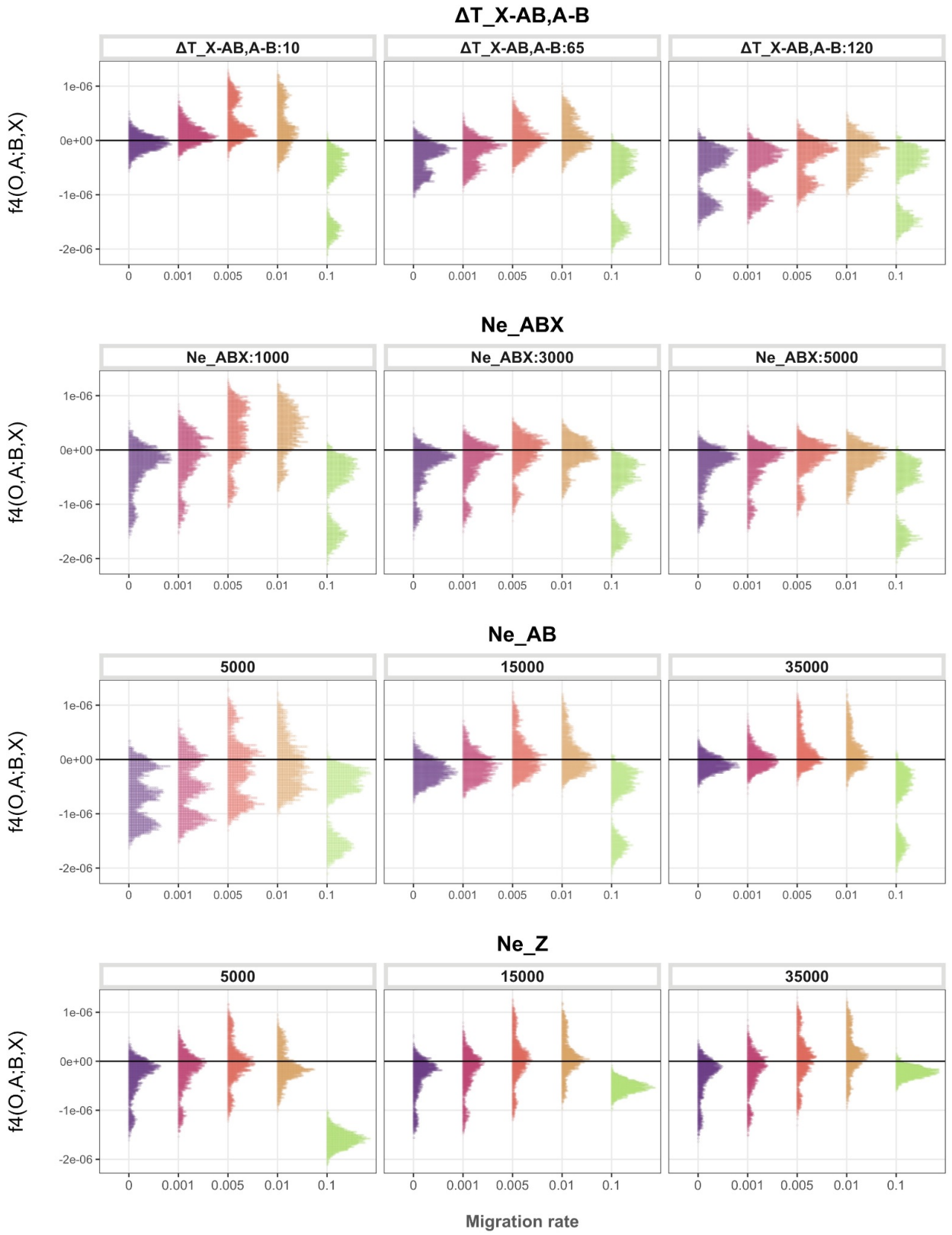


Figure 4: Change in f_4 statistics of the form $f_4(O, A; B, X)$ depending on the varying parameters in population genetic simulations, where negative values represent higher

affinity between A and B, and positive values higher affinity between A and X. The panels for each parameter contain all values of the remaining studied parameters and are explained in Table 1. The multimodal distribution shapes arise because each distribution is collected from a range of simulations run under a range of parameters with diverse effects on the f_4 statistic.

We note several conditions driving more positive $f_4(O, A; B, X)$ values under low-level migration in our model (Figures 4,5; Table 1).

- Low $\Delta T_{X\text{-}AB,A\text{-}B}$ (Figure 4 first row): this means A and B share less drift relative to X (the phylogeny is more star-like). Also, because $T_{AB\text{-}X}$ is fixed in the model, low $\Delta T_{X\text{-}AB,A\text{-}B}$ leads to $T_{A\text{-}B}$ going back in time (Figure 2) and therefore A and B each being impacted by Z-sourced continuous migration for additional generations (i.e. larger α in Figure 3).
- Low Ne_{ABX}: this causes higher shared drift between A, B and X (i.e. larger *g* in Figure 3).
- High Ne_{AB}: this leads to lower shared drift between A and B (i.e. smaller *e* in Figure 3).
- High Ne_{Z}: this causes Z to be subject to less drift, such that the negative branch length *f* in Figure 3 becomes smaller in magnitude.

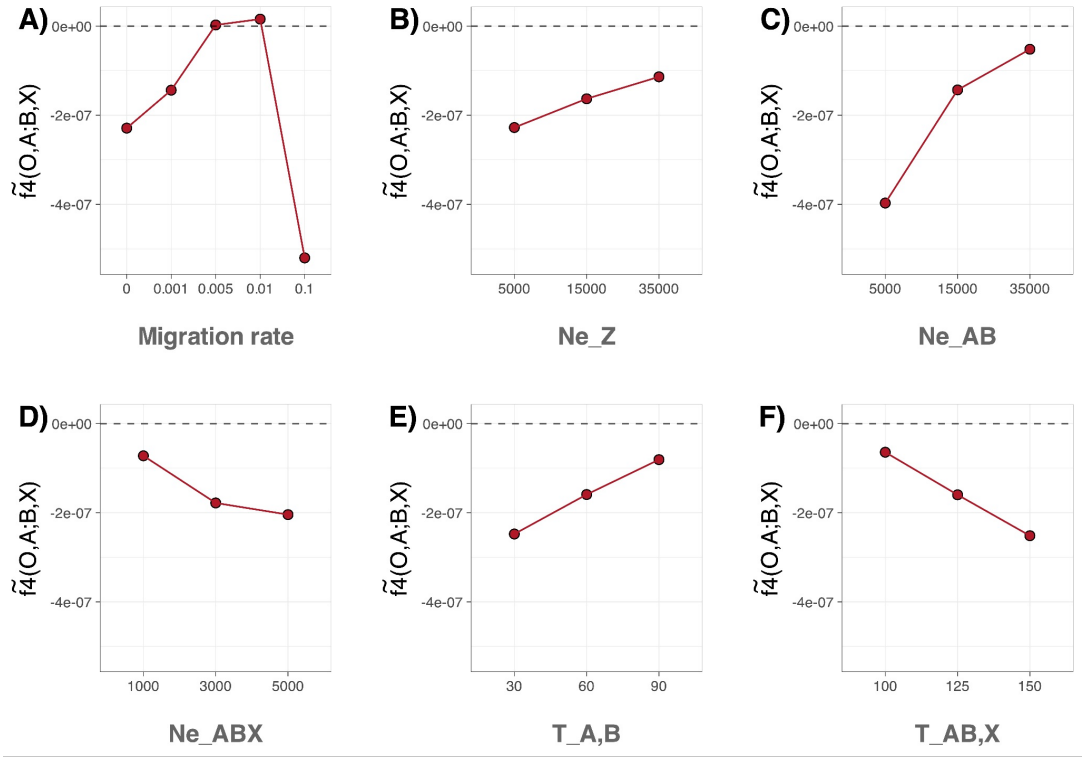


Figure 5: Change in f_4 statistics of the form $f_4(O, A; B, X)$ in population genetic simulations, where all parameters except the one in question were fixed at their median values. The median f_4 value for all replicates is plotted per variable.

As opposed to results involving low-level migration (m=0.001 to m=0.01), at m=0.1, $f_4(O, A; B, X)$ values shift back to negative values, i.e. sister attraction (Figure 4,5, S4). This happens because high migration rates from Z lead to the homogenization of A and B with Z. In other words, A and B become part of the Z gene pool. The drop in f_4 values is particularly strong if Ne_Z is low: e.g. at m=0.1 and Ne_Z = 5k all results are negative (Figure 4,5, S4). The theoretical model in Figure 3 also predicts a similar behaviour when migration probabilities are high.

We then asked whether, in simulations where we detect sister repulsion between A and B in f_4-statistics, we would still observe A and B clustering separately from X in the MDS analysis, a pattern we had observed in the empirical data (Figure 1). For this, we chose one random case from each scenario and calculated outgroup-f_3 statistics. We then compared the mean distance between X to A and B [$\underline{d}_{mds}$(X,A/B)] versus the distance between A and B [d_{mds}(A,B)] in MDS space. In 99% of all cases, populations A and B

clustered separately from X, i.e. d_{mds}(A,B) was lower. This included scenarios where 100% of the f_4 values signified sister repulsion (Figure 6A,B). Meanwhile, under low-level migration we also observed $f_3(O; A, B) < f_3(O; A/B, X)$, reflecting sister repulsion (Figure 6C). Hence, similar to our observation with the East Mediterranean Bronze Age genomes (Figure 1B,D), MDS-based clusterings in our simulations contradict pairwise f_3 statistics.

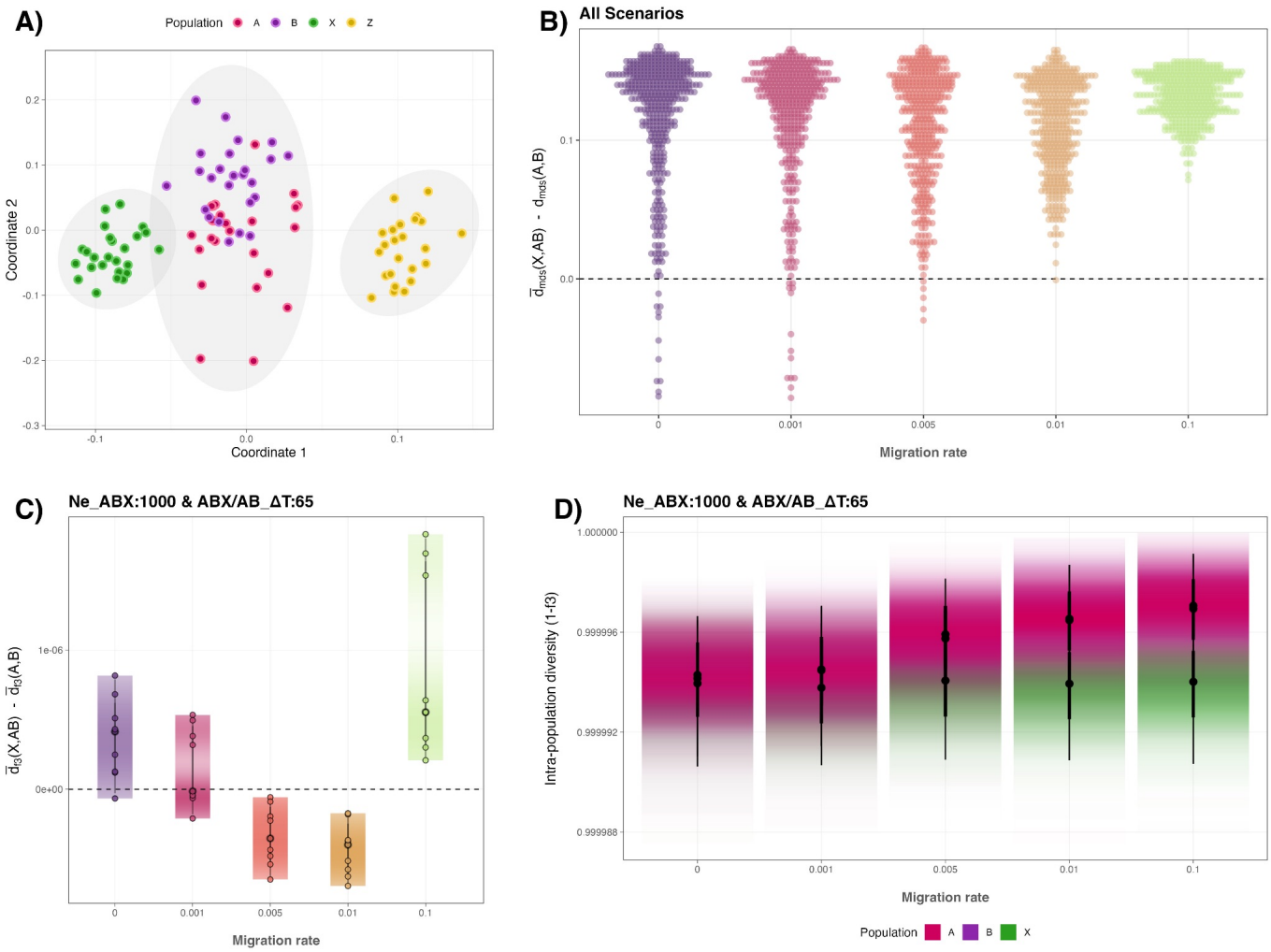


Figure 6: **(A)** MDS of simulated populations A, B, X and Z, using 1 - outgroup-f_3 values as genetic distances, for one sample scenario where $m_{Z\text{->}A,B}$: 0.01 ,Ne_Z : 15000, Ne_{AB} : 15000 , Ne_{ABX} : 1000, and $\Delta T_{X\text{-}AB,A\text{-}B}$: 65. **(B)** MDS clusterings between A and B versus X and A,B, calculated by the Euclidean distances between the mean MDS coordinates of the corresponding populations, for all scenarios. $\underline{d}_{mds}$(X,AB) is the mean distance of X to

A and B. d_{mds}(A,B) is the distance between A and B. Positive values for $\underline{d}_{mds}$(X,AB) - d_{mds}(A,B) represent higher distance of X to A and B, thus distinct clustering of X from A and B. **(C)** Change in f3 values with migration rate from Z for scenarios where Ne_{ABX} : 1000 and $\Delta T_{X\text{-}AB,A\text{-}B}$: 65. $\underline{d}_{f3}$(X,A/B) is the mean 1-f3 distance of X to A or to B. $\underline{d}_{f3}$(A,B) is the mean 1-f_3 distance between A and B. Negative values for $\underline{d}_{f3}$(X,A/B) - $\underline{d}_{f3}$(A,B) show higher affinity between A or B and X than between A and B. **(D)** Intra-population diversity differences between populations A,B and X in relation to migration rate from Z, for scenarios where Ne_{ABX} : 1000 and $\Delta T_{X\text{-}AB,A\text{-}B}$: 65. The diversities are measured by pairwise 1- outgroup f_3 values per population. At migration rates 0 and 0.001, the values for X have higher deviation and overlap with A and B, thus have lower visibility.

In these models, we might expect population differences in within-population diversity, which could be a source of information for inferring unknown admixture history. Indeed, in the Bronze Age dataset we had observed significantly higher diversity in genomes from Anatolia versus modern-day Greece (Figure 1C). In our simulations that involved gene flow from Z, we likewise observed elevated intra-population diversity in A and B versus X (Figure 6D).

We next studied sister repulsion under additional models and assumptions:

a) We asked how asymmetric Z-sourced gene flow, where only population B receives migrants from Z, would impact f_4 statistics. In this case, as might be expected, $f_4(O, A; B, X)$ tends to be positive, i.e. A shows sister repulsion. Conversely, $f_4(O, B; A, X)$ tends to be negative, i.e. B is still attracted to its sister A (Figure S5). This may also be anticipated from Figure 3: if α_B=0, F_4 would always be negative.
b) We tested gene flow from Z into AB, the ancestor population of A and B. This yielded similar patterns to those observed with gene flow to A and B separately, in terms of both f_4, MDS and diversity estimates (Figure S6).
c) Our main simulations comprise continuous gene flow from Z starting from the A-B split. We hence used single pulse migration models, and replicated our findings with continuous gene flow. As expected, these required higher levels of per generation gene flow to produce the same effect (Figure S7); to be compared with Figure 4).
d) Because the values used for the $T_{A\text{-}B}$ parameter (30-90 generations) are higher than what might be anticipated for the split time of two East Anatolian populations, we tested shorter split times, of 5 and 10 generations. We could again replicate sister repulsion patterns whilst at higher rates of migration (Figure S8; to be compared with Figure 4).

e) We asked whether applying a more realistic setting for genomic data simulation would change the results. We repeated the simulations using the HapMap chromosome 1 recombination map (The International HapMap Consortium) and found that the results remained the same (Figure S9).

These results reveal how low-level gene flow from an external source (Z) can lead to sister repulsion, i.e. closely related pairs of populations (A and B) showing higher affinity to a lineage that has split earlier from them (X) in f_3 and f_4 statistics. This effect happens because the f-statistics are point estimates that represent the average across diverse phylogenetic paths in a population's demographic history. In our model, the effect depends on the distance of Z to XAB; therefore the introduction of so-called "deep ancestry" can readily shift the expected values. This also resembles the "ghost admixture" phenomenon when gene flow from distant sources can bias phylogenetic inference (Rogers and Bohlender 2015; Tricou *et al.* 2022).

Finally, we tested whether gene flow from X into A and into B independently (after the A-B split) might also lead to sister repulsion, without invoking the external source Z. We confirmed this was the case using both f_3- and f_4-statistics (Figure S10A). We further found that patterns observed both in the above simulations with external gene flow (Figure 4-6) and in the empirical data (Figure 1), namely lower diversity in X versus A/B and separate MDS clustering of X and A/B, can be replicated under gene flow from X, but only when the effective population size of X (Ne_X) is low relative to Ne_A and Ne_B (Figure S10B,C).

Sister repulsion is an interesting case because in all the scenarios studied, A and B share more similar demographic histories with each other than with X, although this is not directly reflected in the f_3- and f_4-statistics. When sister repulsion is suspected, researchers may resort to alternative tools such as genetic clustering (e.g. MDS), qpAdm modelling (Harney *et al.* 2021) or using time stratifying variants (Speidel *et al.* 2024) to disentangle the signals. As a quick diagnostic of the problem, we highlight the method of randomly splitting one of the target populations, whereby if f_4 statistics show the repulsion signal between the two subgroups, sister repulsion is evident. Researchers may further be interested in evaluating whether post-split admixture from X or an external population Z is more likely, which may be evaluated using additional sources of information, such as local ancestry inference or archaeological/historical evidence for regional gene flow. In the empirical cases we covered, such as sister repulsion in Bronze Age East Anatolia, external gene flow from genetically distant populations (e.g. related to sub-Saharan Africa) may be a more parsimonious explanation than direct gene flow from the population being compared (e.g., Greece).

Methods

Ancient genomes

We used 79 published Bronze Age genomes from the East Mediterranean (Supplementary Table 1), from a publicly available dataset comprising the ancient individuals analyzed in Lazaridis et al., 2022, genotyped at the 1240K array positions (Lazaridis *et al.* 2022a). We only used Bronze Age genomes from East Anatolia and Greece produced with the 1240K enrichment method in order to avoid any technical bias that can arise when co-analyzing different data types, e.g. shotgun versus capture (Koptekin *et al.* 2023).

Principal Components Analysis (PCA)

For the principal components analysis, we used the Human Origins Dataset (Lazaridis *et al.* 2014, 2016), containing 2,583 present-day genomes genotyped at the Affymetrix Human Origins Array positions, merged with the Bronze Age genomes. PCA was performed using *smartpca v.18140*, with option *lsqproject:YES*, implemented in *EIGENSOFT v.7.2.0* (Patterson *et al.* 2006). We projected the ancient individuals onto the first two principal components calculated from present-day Western Eurasian diversity based on populations from the Human Origins panel.

f_3 , f_4 - statistics and Multidimensional scaling (MDS)

We calculated outgroup f_3- and f_4-statistics using *admixr v0.9.1,* which runs *AdmixTools v.7.0.2,* with default parameters (Patterson *et al.* 2012; Petr *et al.* 2019b). Within-group genetic diversities were estimated by computing pairwise 1 - outgroup-f_3. All statistics were computed based on the 1240K dataset, using 3 modern-day Yoruba individuals as the outgroup. We estimated pairwise genetic distances with 1- outgroup f_3, computed a distance matrix of 1-f_3 values, and performed a classical multidimensional scaling analysis using the *cmdscale* algorithm implemented in the R package *stats* (R Core Team, 2023).

The geodesic distances between sites for the comparison of genetic versus geographic distances were calculated using the R package *geosphere* (Hijmans 2010).

For the calculation of f_4(Yoruba, Group1; Group2, Greece) statistics within the East Anatolian site Alalakh, we randomly split the 25 individuals in two groups of 20 & 5 and 15 & 10. Repeating each of the two groupings five times, we computed ten f_4 statistics in total.

Simulations

We used *msprime v1.0* (Baumdicker *et al.* 2022) to simulate diploid sequences of 100Mb, where we sampled 25 individuals from each population. The divergence time of the population used as outgroup (ABXZO) was given as 200 generations ago, followed by the divergence of ABXZ after 20 generations. The effective population size of each population other than the ones given in Table 1 were fixed as 5k, i.e. populations ABXZO, O, ABXZ, X (except for scenarios where Ne_{ABX} = 1k and $\Delta T_{X\text{-}AB,A\text{-}B}$ = 65), A, and B. The mutation rate and the recombination rate were chosen as 1e-8 and 1.1e-8 per generation, respectively (Baumdicker *et al.* 2022). Each scenario had 100 replicates. One-way migration rates from Z were set the same into both A and B, except for one case (Ne_Z, Ne_{AB}, Ne_{ABX}, $\Delta T_{X\text{-}AB,A\text{-}B}$) where we tested the effect of asymmetric migration by allowing gene flow (m=0.01) only into B. For scenarios where Ne_{ABX} = 1k and $\Delta T_{X\text{-}AB,A\text{-}B}$ = 65, one-way migration from X into A and B, migration from Z only into AB and single pulse migration models at times $T_{A\text{-}B}/2$ and $T_{AB\text{-}X}/2$ with additional migration rates (0.15 and 0.2) were also tested. All combinations of values for parameters $m_{Z\text{->}A.B}$, Ne_Z, Ne_{AB}, Ne_{ABX}, and $\Delta T_{X\text{-}AB,A\text{-}B}$ given in Table 1 were tested, in addition to scenarios where $T_{A\text{-}B}$ was set to 5 and 10 generations ago. We also ran simulations using the HapMap recombination map for chromosome 1 (The International HapMap Consortium), under scenarios where Ne_{ABX} = 1k and $\Delta T_{X\text{-}AB,A\text{-}B}$ = 65. The outgroup-f_3 and f_4 statistics were calculated using *tskit v0.5.6*, which follows the calculation in Patterson et al., 2012, but the default *span_normalise=True* option we used normalizes the values by the total length of windows, which results in differences in the scale of the f_3 and f_4 values, *tskit* yielding smaller values than Admixtools. The approach for split-f_4 analysis explained above for real data was followed for a randomly chosen scenario where Ne_{ABX} = 1k, $\Delta T_{X\text{-}AB,A\text{-}B}$ = 65 and m=0.01. The f_4(O, Group1; Group2,B) were calculated between groups of population A. The MDS analysis was performed the same way as above. In order to compare the MDS distances between populations, we used the mean MDS x and y coordinates of each individual, and calculated the mean coordinate point per population. We then calculated Euclidean distances between the mean coordinates of the corresponding populations.

Data Availability

The code and parameters for generating the simulated data are deposited under https://github.com/goztag/msprime_fstats/tree/main.

Acknowledgements

We thank Emilia Huerta-Sanchez, Torsten Günther, Benjamin Vernot, Pavlos Pavlidis, Dilek Koptekin, Yetkin Alıcı, Iosif Lazaridis and members of the METU CompEvo Team and the NEOMATRIX Collective for helpful discussions. The work was supported by the H2020 ERC Consolidator grant (no. 772390 NEOGENE to M.S.), H2020-WIDESPREAD-05-2020 TWINNING grant (no. 952317 NEOMATRIX to M.S.) and Scientific and Technological Research Council of Turkey (TÜBİTAK) through the 2210/A National Scholarship Programme for MSc Students (to G.A.).

References

Antonio M. L., C. L. Weiß, Z. Gao, S. Sawyer, V. Oberreiter, et al., 2024 Stable population structure in Europe since the Iron Age, despite high mobility. eLife 13. https://doi.org/10.7554/elife.79714

Altınışık N. E., D. D. Kazancı, A. Aydoğan, H. C. Gemici, Ö. D. Erdal, et al., 2022 A genomic snapshot of demographic and cultural dynamism in Upper Mesopotamia during the Neolithic Transition. Science Advances 8. https://doi.org/10.1126/sciadv.abo3609

Baumdicker F., G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, *et al.*, 2022 Efficient ancestry and mutation simulation with msprime 1.0. Genetics 220: iyab229. https://doi.org/10.1093/genetics/iyab229

Bergström A., S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, *et al.*, 2020 Insights into human genetic variation and population history from 929 diverse genomes. Science 367: eaay5012. https://doi.org/10.1126/science.aay5012

Clemente F., M. Unterländer, O. Dolgova, C. E. G. Amorim, F. Coroado-Santos, *et al.*, 2021 The genomic history of the Aegean palatial civilizations. Cell 184: 2565-2586.e21. https://doi.org/10.1016/j.cell.2021.03.039

Durand E. Y., N. Patterson, D. Reich, and M. Slatkin, 2011 Testing for ancient admixture between closely related populations. Mol Biol Evol 28: 2239–2252. https://doi.org/10.1093/molbev/msr048

Eriksson A., and A. Manica, 2012 Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proceedings of the National Academy of Sciences of the United States of America 109: 13956–13960. https://doi.org/10.1073/pnas.1200567109

Flegontov P., U. Işıldak, R. Maier, E. Yüncü, P. Changmai, *et al.*, 2023 Modeling of African population history using f-statistics is biased when applying all previously proposed SNP

ascertainment schemes. PLOS Genetics 19: e1010931. https://doi.org/10.1371/journal.pgen.1010931

Fu Q., C. Posth, M. Hajdinjak, M. Petr, S. Mallick, et al., 2016 The genetic history of Ice Age Europe. Nature 534: 200–205. https://doi.org/10.1038/nature17993

The International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861. https://doi.org/10.1038/nature06258

Harney É., N. Patterson, D. Reich, and J. Wakeley, 2021 Assessing the performance of qpAdm: a statistical tool for studying population admixture. Genetics 217. https://doi.org/10.1093/genetics/iyaa045

Harris A. M., and M. DeGiorgio, 2017 Admixture and Ancestry Inference from Ancient and Modern Samples through Measures of Population Genetic Drift. Hum Biol 89: 21–46. https://doi.org/10.13110/humanbiology.89.1.02

Kılınç G. M., D. Koptekin, Ç. Atakuman, A. P. Sümer, H. M. Dönertaş, *et al.*, 2017 Archaeogenomic analysis of the first steps of Neolithization in Anatolia and the Aegean. Proceedings of the Royal Society B: Biological Sciences 284: 20172064. https://doi.org/10.1098/rspb.2017.2064

Koptekin D., E. Yapar, K. B. Vural, E. Sağlıcan, N. E. Altınışık, et al., 2023 Pre-processing of paleogenomes: Mitigating reference bias and postmortem damage in ancient genome data. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2023.11.11.566695

Kouka O., 2016 The Built Environment and Cultural Connectivity in the Aegean Early Bronze Age, in Molloy, B.P.C (ed.), Of Odysseys and Oddities, Sheffield Studies in Aegean Archaeology, Oxbow Books, 203-222.

Lazaridis I., N. Patterson, A. Mittnik, G. Renaud, S. Mallick, *et al.*, 2014 Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513: 409–413. https://doi.org/10.1038/nature13673

Lazaridis I., D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, *et al.*, 2016 Genomic insights into the origin of farming in the ancient Near East. Nature 536: 419–424. https://doi.org/10.1038/nature19310

Lazaridis I., A. Mittnik, N. Patterson, S. Mallick, N. Rohland, *et al.*, 2017 Genetic origins of the Minoans and Mycenaeans. Nature 548: 214–218. https://doi.org/10.1038/nature23310

Lazaridis I., S. Alpaslan-Roodenberg, R. Pinhasi, and D. Reich, 2022a The genetic history of the Southern Arc: A bridge between West Asia and Europe

Lazaridis I., S. Alpaslan-Roodenberg, A. Acar, A. Açıkkol, A. Agelarakis, *et al.*, 2022b The genetic history of the Southern Arc: A bridge between West Asia and Europe. Science 377: eabm4247. https://doi.org/10.1126/science.abm4247

Lipson M., 2020 Applying f4-statistics and admixture graphs: Theory and examples. Molecular Ecology Resources 20: 1658–1667. https://doi.org/10.1111/1755-0998.13230

Margaryan A., D. J. Lawson, M. Sikora, F. Racimo, S. Rasmussen, et al., 2020 Population genomics of the Viking world. Nature 585: 390–396. https://doi.org/10.1038/s41586-020-2688-8

Massa M., 2016 Networks before Empires: cultural transfers in western and central Anatolia during the Early Bronze Age

Patterson N., A. L. Price, and D. Reich, 2006 Population Structure and Eigenanalysis. PLOS Genetics 2: e190. https://doi.org/10.1371/journal.pgen.0020190

Patterson N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, *et al.*, 2012 Ancient Admixture in Human History. Genetics 192: 1065–1093. https://doi.org/10.1534/genetics.112.145037

Peter B. M., 2016 Admixture, Population Structure, and F-Statistics. Genetics 202: 1485–1501. https://doi.org/10.1534/genetics.115.183913

Peter B. M., 2022 A geometric relationship of F 2 , F 3 and F 4 -statistics with principal component analysis. Philosophical Transactions - Royal Society. Biological Sciences 377. https://doi.org/10.1098/rstb.2020.0413

Petr M., S. Pääbo, J. Kelso, and B. Vernot, 2019a Limits of long-term selection against Neandertal introgression. Proc. Natl. Acad. Sci. U.S.A. 116: 1639–1644. https://doi.org/10.1073/pnas.1814338116

Petr M., B. Vernot, and J. Kelso, 2019b admixr—R package for reproducible analyses using ADMIXTOOLS. Bioinformatics 35: 3194–3195. https://doi.org/10.1093/bioinformatics/btz030

R Core Team, 2023 _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Hijmans R. J. (2010). geosphere: Spherical Trigonometry. R package version 1.5-18, https://cran.r-project.org/web/packages/geosphere

Rodríguez-Varela R., T. Günther, M. Krzewińska, J. Storå, T. H. Gillingwater, *et al.*, 2017 Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. Current Biology 27: 3396-3402.e5. https://doi.org/10.1016/j.cub.2017.09.059

Rogers A. R., and R. J. Bohlender, 2015 Bias in estimators of archaic admixture. Theoretical Population Biology 100: 63–78. https://doi.org/10.1016/j.tpb.2014.12.006

Rohland N., S. Mallick, M. Mah, R. Maier, N. Patterson, et al., 2022 Three assays for in-solution enrichment of ancient human DNA at more than a million SNPs. Genome Research 32: 2068–2078. https://doi.org/10.1101/gr.276728.122

Speidel L., M. Silva, T. Booth, B. Raffield, K. Anastasiadou, et al., 2024 High-resolution genomic ancestry reveals mobility in early medieval Europe. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2024.03.15.585102

Şahoğlu V., 2005 The Anatolian Trade Network and the Izmir region during the Early Bronze Age. Oxford Journal of Archaeology 24: 339–361. https://doi.org/10.1111/j.1468-0092.2005.00240.x

Skourtanioti E., Y. S. Erdal, M. Frangipane, F. Balossi Restelli, K. A. Yener, *et al.*, 2020 Genomic History of Neolithic to Bronze Age Anatolia, Northern Levant, and Southern Caucasus. Cell 181: 1158-1175.e28. https://doi.org/10.1016/j.cell.2020.04.044

Skourtanioti E., H. Ringbauer, G. A. Gnecchi Ruscone, R. A. Bianco, M. Burri, *et al.*, 2023 Ancient DNA reveals admixture history and endogamy in the prehistoric Aegean. Nat Ecol Evol 1–14. https://doi.org/10.1038/s41559-022-01952-3

Tricou T., E. Tannier, and D. M. de Vienne, 2022 Ghost Lineages Highly Influence the Interpretation of Introgression Tests. Systematic Biology 71: 1147–1158. https://doi.org/10.1093/sysbio/syac011