



HAL
open science

Inference of population genetic structure from temporal samples of DNA

Olivier François, Séverine Liégeois, Benjamin Demaille, Flora Jay

► **To cite this version:**

Olivier François, Séverine Liégeois, Benjamin Demaille, Flora Jay. Inference of population genetic structure from temporal samples of DNA. 2019. hal-02413974

HAL Id: hal-02413974

<https://hal.science/hal-02413974v1>

Preprint submitted on 16 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Inference of population genetic structure from temporal samples of DNA

Olivier François^{1,†} Séverine Liégeois² Benjamin Demaille²
Flora Jay^{2,†}

Authors' affiliations.

¹ Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Grenoble INP, Laboratoire TIMC-IMAG UMR 5525, 38000 Grenoble, France.

² Université Paris Sud, Laboratoire de Recherche en Informatique, Centre National de la Recherche Scientifique, Bâtiment 650 Ada Lovelace, 91405 Orsay Cedex, France.

† Correspondence to: flora.jay@lri.fr, olivier.francois@imag.fr.

1

Abstract

2 The recent years have seen a growing number of studies investigating evolutionary
3 questions using ancient DNA techniques and temporal samples of DNA. To address
4 these questions, one of the most frequently-used algorithm is based on principal com-
5 ponent analysis (PCA). When PCA is applied to temporal samples, the sample dates
6 are, however, ignored during analysis, which could lead to some misinterpretations of
7 the results. Here we introduce a new factor analysis (FA) method for which individ-
8 ual scores are corrected for the effect of allele frequency drift through time. Based
9 on a diffusion approximation, our approach approximates allele frequency drift in
10 a random mating population by a Brownian process. Exact solutions for estimates
11 of corrected factors are obtained, and a fast estimation algorithm is presented. We
12 compared data representations obtained from the FA method with PCA and with PC
13 projections in simulations of divergence and admixture scenarios. Then we applied
14 FA with correction for temporal drift to study the evolution of hepatitis C virus in
15 a patient infected by multiple strains, and to describe the population structure of
16 ancient European samples.

17 **1 Introduction**

18 In recent years, the number of studies analyzing temporal samples of DNA or ancient
19 DNA has increased dramatically, both for humans and for other organisms (Lazaridis
20 *et al.*, 2014; Haak *et al.*, 2015; Mathieson *et al.*, 2015; Carroll *et al.*, 2015; Skoglund
21 and Mathieson, 2018). In such studies, a central question concerns the inference of
22 ancestral relationships between sampled populations (Slatkin, 2016). Evolutionary
23 biologists and population geneticists have devised many methods for addressing this
24 question. One of the most frequently-used method is based on principal component
25 analysis (PCA) and projections of ancient samples on axes built from present-day
26 samples (Patterson *et al.*, 2006, 2012). In population genetics, PCA is performed
27 by finding the eigenvalues and eigenvectors, or axes, of the covariance matrix of al-
28 lele frequencies. The highest order eigenvectors indicate the directions in the high
29 dimensional allele-frequency space which account for most of the covariance. Indi-
30 vidual samples are then plotted on the plane spanned by the first axes, offering a
31 visual representation of the structure hidden in the data obtained with short com-
32 puting time. Relative distances in the reduced space indicate their similarity and
33 their ancestral relationships (McVean, 2009). When PCA or PC projections are ap-
34 plied to analyze temporal samples, information on sample dates is, however, usually
35 omitted in the computation of eigenvalues and eigenvectors (Slatkin, 2016; Slatkin
36 and Racimo, 2016; Harris and DeGiorgio, 2017).

37 Previous studies have reported that time differences in samples are reflected in the
38 principal axes of a PCA (Skoglund *et al.*, 2014), creating sinusoidal shapes similar to
39 those observed with geographic samples (Novembre and Stephens, 2008). The combi-
40 nation of both time and spatial heterogeneity in sampling further modify the patterns

41 observed in PCA. Local dispersal through time causes ancient samples to be shrunk
42 toward the center of the PC plot and not to cluster with their present-day counter-
43 part despite no major discontinuity in the demographic process (Duforet-Frebourg
44 and Slatkin, 2016). Sinusoidal distortions linked to gradients and longitudinal data
45 also occur in various fields, and are called *horseshoes* or *arches*. Those distortions
46 complicate the interpretation of multidimensional scaling, local kernel methods and
47 ordination analysis (Hill and Gauch, 1980; Diaconis *et al.*, 2008). Supervised methods
48 that combine ancient and modern samples by using PC projections on present-day
49 samples also suffer from some statistical issues. PC projections exhibit a shrinkage
50 bias toward the center of the principal axes, and this bias could increase in analyses
51 of temporal samples (Lee *et al.*, 2010). Since those biases could lead to misinterpre-
52 tations or to incorrect estimates of individual ancestry, it is important to propose
53 methods that correct principal components when temporal samples are analyzed for
54 descriptive purposes.

55 Corrections of sinusoidal patterns arising in principal components have been pro-
56 posed when distortions are caused by spatial auto-correlation in geographic samples
57 (Frichot *et al.*, 2012). Similarly, Kalaitzis and Lawrence (2012) have proposed to
58 remove temporal correlations leaving residual variance with residual component anal-
59 ysis. Modified versions of the STRUCTURE algorithm – which is closely related to
60 PCA – were also developed to integrate corrections based on spatial or temporal diffu-
61 sion models (Pritchard *et al.*, 2000; Caye *et al.*, 2018; Joseph and Pe’er, 2018). In this
62 study, we introduce a new factor analysis (FA) method for visualizing hidden struc-
63 ture and for describing ancestral relationships among samples collected at distinct
64 time points in the past. Based on a diffusion approximation, our approach approxi-

65 mates allele frequency drift in a random mating population by a Brownian process.
66 Using the Karhunen-Loève theorem, we propose a representation of the factor model
67 in which additional covariates, representing temporal eigenvectors, are introduced
68 in the model. Our model assumes informative Gaussian prior distributions for the
69 effect sizes of the temporal covariates. Exact solutions for time-corrected factors
70 are obtained, and a fast algorithm based on singular value decomposition (SVD) is
71 proposed. We compare corrections for temporal drift in FA with PCA in coalescent
72 and generative simulations of divergence and admixture scenarios. We eventually
73 apply corrections for temporal drift to study the evolution of hepatitis C virus in a
74 patient infected by multiple viral strains, and to describe population structure for
75 DNA samples from ancient Europeans and Eurasians.

76 **2 New Method**

77 This section introduces a new factor analysis method for describing ancestry among
78 samples taken at distinct time points in the past. The objective is to propose a
79 factorial decomposition of the data matrix similar to a PCA, in which the individual
80 scores are corrected for the effect of allele frequency drift through time. The scores,
81 called factors, will be obtained as maximum-a-posteriori estimates in a Bayesian
82 model.

83 **Model.** Let \mathbf{Y} be an $n \times p$ matrix of genotypic data, where n is the number of
84 individual samples and p is the number of markers, typically represented as single
85 nucleotide polymorphisms (SNPs). We suppose that the data are centered, so that
86 the mean value for each column (or marker) is null. We also suppose that each sample,

87 i , is associated with a sampling date, t_i , corresponding to the age of the sample. The
88 dates are normalized to span the unit interval $0 < t_1 \leq \dots \leq t_n \leq 1$. Here, time has
89 a forward representation. The date t_1 corresponds to the most ancient sample, and
90 $t = 1$ represents samples at present time. Our FA model takes the following form

$$\mathbf{Y} = \mathbf{UV}^T + \epsilon, \quad (1)$$

91 where \mathbf{U} is an $n \times K$ matrix of scores, \mathbf{V}^T is a $K \times p$ matrix of loadings. The
92 number of factors, K , can be set to any number smaller than n and p depending
93 on how drastically one wants to reduce the dimension of the data (and approximate
94 the data matrix). It can be set to the number of ancestral groups minus one when
95 this information is known. The individual scores contained in the K column vectors,
96 $\mathbf{u}_1, \dots, \mathbf{u}_K$, of the matrix \mathbf{U} reflect the ancestral relationships among samples (Pat-
97 terson *et al.*, 2006). To incorporate corrections for temporal drift, we model the error
98 term, ϵ , as follows

$$\epsilon \sim \mathbf{N}(0, \alpha^{-1}\mathbf{C} + \sigma^2\mathbf{I}), \quad (2)$$

99 where $\mathbf{N}(0, \alpha^{-1}\mathbf{C} + \sigma^2\mathbf{I})$ is the multidimensional Gaussian distribution with mean 0
100 and covariance matrix $\alpha^{-1}\mathbf{C} + \sigma^2\mathbf{I}$, α is a precision (scale) parameter for temporal
101 drift, σ^2 is the variance of the residual error, and \mathbf{I} is the $n \times n$ identity matrix. We
102 suppose that \mathbf{C} is an $n \times n$ covariance matrix given by

$$c_{ij} = \min(t_i, t_j) \quad i, j = 1, \dots, n. \quad (3)$$

103 The definition of the covariance matrix, \mathbf{C} , is related to the covariance function
104 of the Brownian process. This model assumption corresponds to the diffusion ap-
105 proximation of allele frequency drift in a random mating population conditional on

106 non-fixation of alleles in the population (Kimura, 1964, 1983). The diffusion approx-
107 imation underlies the development of several recent methods of ancestry estimation
108 similar to our model (Patterson *et al.*, 2012; Pickrell and Pritchard, 2012; Peter, 2016;
109 Joseph and Pe'er, 2018). As a consequence of the definition, the variance of allele
110 frequencies is proportional to time. In applications, we normalized the sample dates
111 so that t_1 corresponds to the variance of allele frequencies in the oldest sample.

112 **Factor estimates.** To compute the factor matrix, \mathbf{U} , in the model equation (1),
113 we turned to an equivalent formulation of this equation

$$\mathbf{Y} = \mathbf{W} + \mathbf{Z}\mathbf{B}^T + \epsilon', \quad (4)$$

114 where the residual noise is described by

$$\epsilon' \sim \mathbf{N}(0, \sigma^2 \mathbf{I}). \quad (5)$$

115 In this formula, effect sizes, B_j , $j = 1, \dots, p$, are introduced, and considered as
116 i.i.d. random variables with univariate Gaussian prior distribution $N(0, \alpha^{-1})$. A
117 latent matrix, $\mathbf{W} = \mathbf{U}\mathbf{V}^T$, has a non-informative prior distribution. After a spectral
118 decomposition of the covariance matrix \mathbf{C} , we define

$$\mathbf{Z} = \mathbf{P}\sqrt{\mathbf{\Lambda}} \quad (6)$$

119 where \mathbf{P} is a unitary matrix of eigenvectors, and $\mathbf{\Lambda}$ is the diagonal matrix containing
120 the eigenvalues of \mathbf{C}

$$\mathbf{C} = \mathbf{Z}\mathbf{Z}^T = (\mathbf{P}\sqrt{\mathbf{\Lambda}})(\mathbf{P}\sqrt{\mathbf{\Lambda}})^T. \quad (7)$$

121 Based on the Karhunen-Loève theorem (Loève, 1948), the diagonal terms of $\mathbf{\Lambda}$ can
122 be approximated as

$$\lambda_i \approx \frac{n}{(i - 1/2)^2 \pi^2}, \quad i = 1, \dots, n, \quad (8)$$

123 and we have

$$Z_{ij} \approx f_i(t_j) \sqrt{\lambda_i/n}, \quad i, j = 1, \dots, n, \quad (9)$$

124 where $f_i(t)$ is defined as $f_i(t) = \sqrt{2} \sin((i - 1/2)\pi t)$ for all t in the interval $[0, 1]$.

125 According to these results, the eigenvectors of the covariance matrix have sinusoidal
126 shapes, and a diffusion model is consistent with the arch effect observed in principal
127 components of genetic variation (Skoglund *et al.*, 2014).

128 Statistical estimates of the matrices \mathbf{U} , \mathbf{V} and \mathbf{B} can be obtained by maximizing
129 a posterior distribution in a Bayesian framework. This approach amounts to finding
130 the minimum of the following loss function

$$\mathcal{L}(\mathbf{W}, \mathbf{B}) = \frac{1}{2} \|\mathbf{Y} - \mathbf{W} - \mathbf{ZB}^T\|_F^2 + \frac{1}{2} \lambda \|\mathbf{B}\|^2, \quad (10)$$

131 where we have set λ equal to the inverse of the temporal signal-to-noise ratio, $\lambda =$
132 $\alpha\sigma^2$. Finding the matrices \mathbf{W} and \mathbf{B} that minimize the loss function $\mathcal{L}(\mathbf{W}, \mathbf{B})$ is
133 equivalent to computing their estimates in a latent factor regression model with ridge
134 penalty (Frichot *et al.*, 2013). According to Caye *et al.* (2019), the latent matrix, \mathbf{W} ,
135 minimizes the following loss function

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2} \|\mathbf{D}_\lambda \mathbf{P}^T (\mathbf{Y} - \mathbf{W})\|_F^2, \quad (11)$$

136 where \mathbf{D}_λ is a diagonal matrix with coefficients equal to

$$\mathbf{D}_\lambda(i, i) = \left(\frac{\lambda}{\lambda + \lambda_i} \right)^{1/2}, \quad i = 1, \dots, n. \quad (12)$$

137 The estimate of \mathbf{W} is provided by the best approximation of rank K of the matrix
138 \mathbf{Y} , where “best approximation” is related to the following matrix norm

$$\|\mathbf{Y}\|_A^2 = \text{Tr}(\mathbf{Y}^T \mathbf{A} \mathbf{Y}), \quad \mathbf{A} = \mathbf{P} \mathbf{D}_\lambda^2 \mathbf{P}^T. \quad (13)$$

139 In closed form, the optimal solution is equal to

$$\mathbf{W} = \mathbf{P}\mathbf{D}_\lambda^{-1}\text{svd}_K(\mathbf{D}_\lambda\mathbf{P}^T\mathbf{Y}). \quad (14)$$

140 The K corrected factors forming \mathbf{U} and their associated loadings, \mathbf{V} , can be obtained
141 from the SVD of the matrix \mathbf{W} (see Table S1). For very large data sets, a modification
142 of the SVD based on random projections could provide an accelerated version of the
143 algorithm (Halko *et al.*, 2011).

144 **Software availability.** A short working R code presenting the algorithm in a self-
145 contained way is provided in Table S1. The method described in this section is
146 currently implemented as an R package `temporalFA`.

147 3 Results

148 **Horseshoe effect.** To provide an example of distortion arising in PCA due to
149 uncorrected temporal drift, we performed a simulation of a coalescent model for
150 forty-one samples with ages ranging from 0 to 4,000 generations in a population
151 with effective size $N_e = 10,000$. The sample dates in the simulation corresponded
152 to an interval of 100 generations. Covariance among samples was smaller for the
153 most ancient samples than for the most recent samples, and it increased linearly
154 with time (Figure 1A). The patterns observed in the sample covariance matrix were
155 highly similar to those obtained in a theoretical covariance function corresponding
156 to a Brownian process (Figure 1C). The PC plots of individual samples exhibited
157 sinusoidal patterns, in which the most ancient and recent samples were placed at
158 both extremes of a horseshoe (Figure 1B). Correcting for temporal drift, the factor

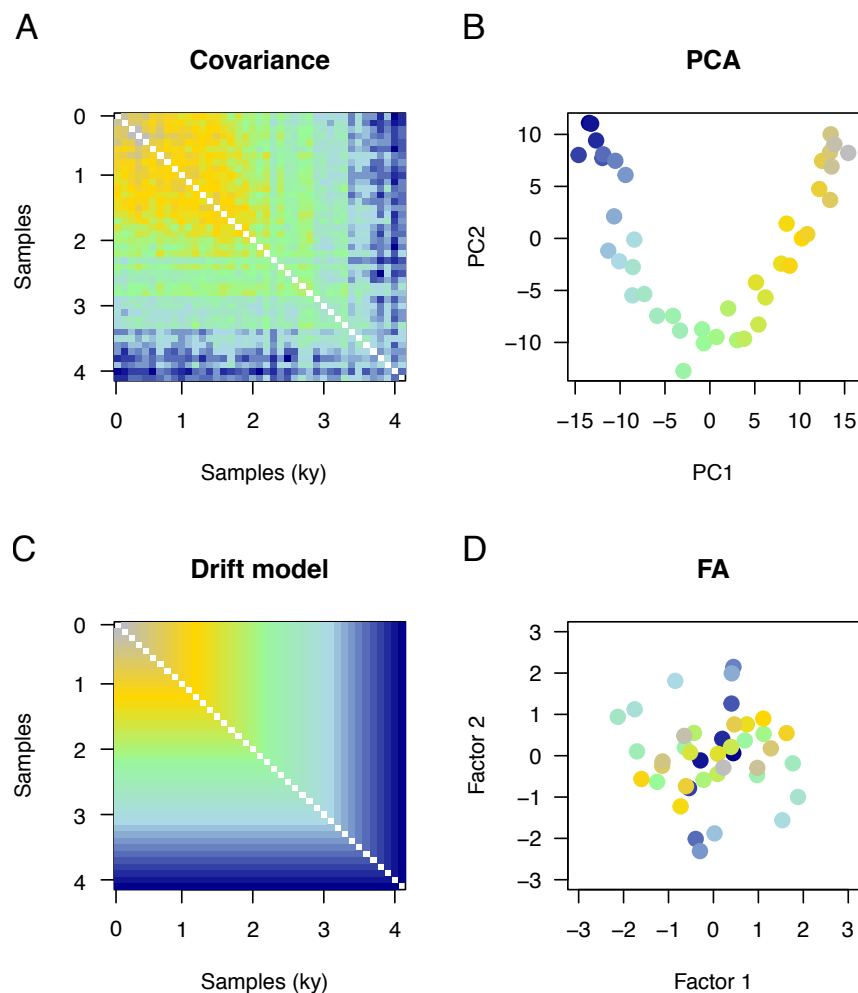


Figure 1. Horseshoe effect. Coalescent simulation of allele frequencies drifting through time in a single population ($N_e = 10,000$). Forty-one samples with ages ranging from 0 (present, grey color) to 4,000 generations (past, dark blue color) were simulated. A) Covariance matrix for observed samples B) PC plot for individual samples, C) Brownian covariance matrix used as a correction model, D) Factor analysis plot showing correction for temporal drift. In covariance matrices, the blue color indicates lower values whereas the yellow and grey colors indicate higher values.

159 analysis plot displayed a single cluster grouping all samples without any apparent
160 structure among samples (Figure 1D). This last result showed that distortion due to
161 temporal drift was correctly removed in a factor analysis using a Brownian model of
162 genetic drift.

163 **Divergence model.** In a second series of experiments, we simulated models of
164 divergence of two populations. In coalescent simulations, twenty-four samples with
165 ages ranging from 0 to 1,000 generations were simulated, corresponding to a sampling
166 interval of 100 generations and four present-day individuals. In a PCA of simulated
167 samples, PC1 reflected the level of divergence between populations while PC2 repre-
168 sented temporal drift (Figure 2A). Correcting for temporal drift, the factor analysis
169 plot exhibited two clusters without any apparent structure within each group (Figure
170 2B). The Davies-Bouldin clustering index reached higher values in the FA plots than
171 in the PC plots, meaning that the clusters were better characterized and better rep-
172 resented populations of origin in FA than in PCA (Figure 3C). In generative model
173 simulations, factor 1 in FA better explained the hidden factor than did the first PC
174 in PCA (Figure 3D). The results provided evidence that correcting for temporal drift
175 in FA revealed population structure hidden in the noisy data.

176 **Admixture models.** In another series of experiments, we considered admixture
177 models in which an ancestral population splits into two sister populations 1,300 gen-
178 erations ago. The two divergent populations came into contact 500 generations ago,
179 giving rise to descendants having 75% ancestry in the first ancestral population and
180 25% ancestry in the second ancestral population. One hundred present-day individ-
181 uals were sampled from the admixed population, and fifty individuals were sampled

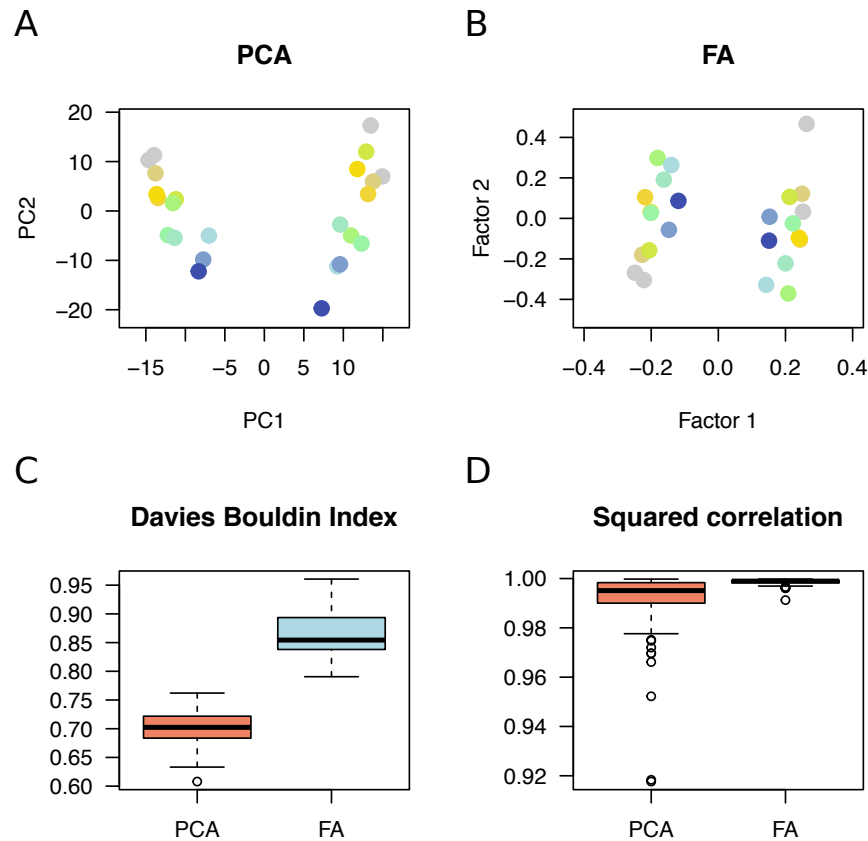


Figure 2. Simulation of two-population models. Twenty-four samples with ages ranging from 0 (present, grey color) to 1,000 generations (past, dark blue color) were simulated. A) Typical PC plot for observed samples, B) Factor analysis plot showing correction for temporal drift, C) Clustering index for PCA and FA results (100 coalescent simulations), D) Squared correlation between PC1 - Factor 1 and a true factor having two modes (100 generative model simulations).

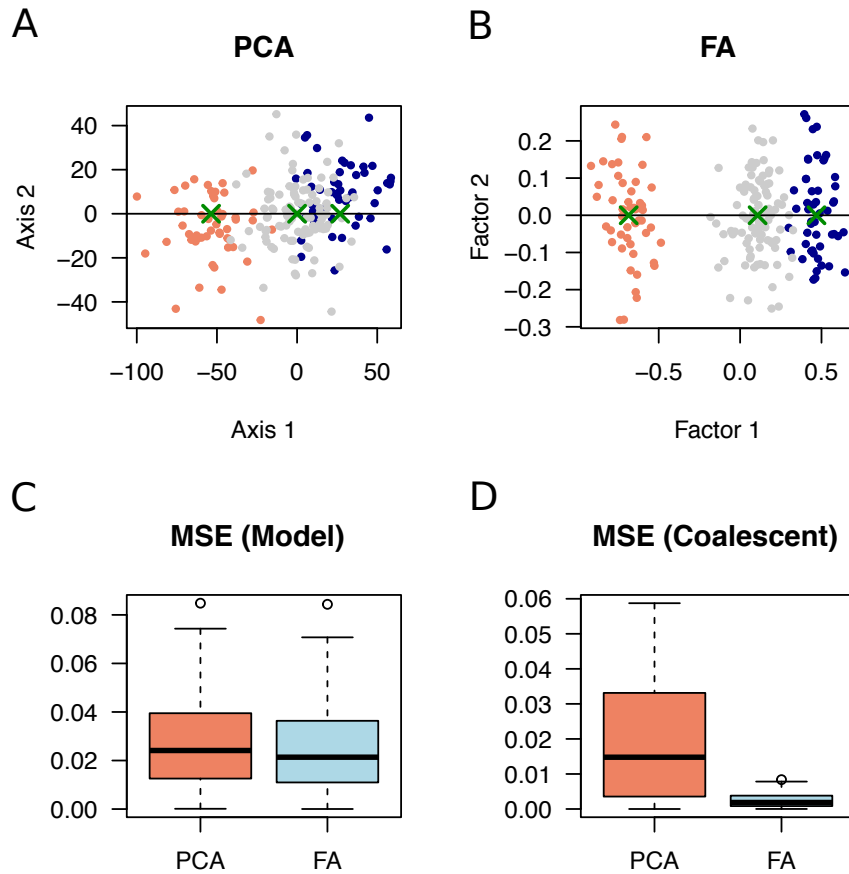


Figure 3. Simulation of admixture models. One hundred samples were simulated for present-day admixed individuals (admixture rate 25 and 75%, grey color) and samples from two ancestral populations (age 1,000 generations, orange and blue colors). A) Typical plot for PC projection of ancient samples onto the admixed population showing a shrinkage effect, B) Factor analysis plot showing correction for shrinkage, C) Mean square error for estimates of admixture proportions from PC projections and FA plots (100 generative model simulations), D) Mean squared error for estimates of admixture proportions (100 coalescent simulations). Green crosses represent population centers, from which admixture estimates were computed.

182 from each ancestral population before the admixture event (1,000 generations ago).
183 Artificial genotypes generated according to a Brownian model were also used to simu-
184 late levels of admixture similar to those observed in coalescent models (see Methods).
185 The objective of the experiments was to compare the results of PC projections of an-
186 cient samples onto the present-day population with those obtained in factor analysis
187 with correction for temporal drift. Typical plots for PC projections exhibited a
188 shrinkage effect in which the projected samples were shifted toward zero, and closer
189 to the admixed population than expected (Figure 3A). The shrinkage effect was even
190 more pronounced in coalescent simulations than in generative model simulations (Fig-
191 ure S1 and Figure 3C-D). Correction for temporal drift in factor analysis removed
192 the shrinkage effect, and, in the FA plot, the locations of centers of ancestral clusters
193 reflected admixture levels more precisely than in PC plots (Figure 3B). The mean
194 squared errors for estimates of admixture proportions were higher in PC projections
195 than in FA plots both in generative and in coalescent simulations (Figure 3C-D). The
196 results showed that correcting for temporal drift in FA improved the representation
197 of admixed individuals and their source populations compared with projections on
198 present-day individual PCs.

199 **Hepatitis C virus infection.** To follow chronic infection in a non-responder hep-
200 atitis C patient treated in the 2000's, we studied $n = 1,934$ samples of viral RNA
201 sequences over a period of thirteen years (Caporossi *et al.*, 2019). The patient was
202 coinfecting by viral strains from two HCV genotypes, 4k and 1b. Height serum sam-
203 ples were available from years 2002 to 2014. Treatment with dual therapy had been
204 administered for six months after the beginning of the follow-up period. A PC plot

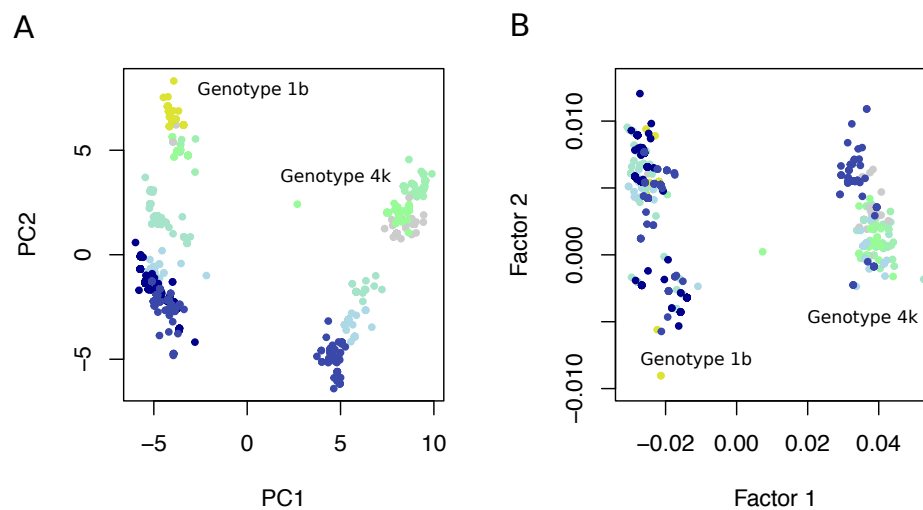


Figure 4. Hepatitis C virus infection. Longitudinal study of a single non-responder patient infected by two viral strains (HCV genotypes 1b and 4k). A total of $n = 1,934$ viral samples were collected from years 2002 to 2014. Dark blue color corresponds to the oldest samples, while yellow and grey colors correspond to the most recent samples. A) PC plot of viral samples, B) factor analysis with correction for temporal drift. A few outlier individuals detected in the FA plot are not shown in the plot.

205 of viral samples displayed a pattern similar to those observed in simulations of di-
206 vergence models (Figure 4A and Figure 2A). PC1 reflected divergence among the
207 samples classified in distinct viral types, and PC2 was influenced by the ages of the
208 samples. After correction for temporal drift in a FA plot, viral particles were grouped
209 according to their phylogenetic classification (Figure 4B). In the FA plot, a first clus-
210 ter consisted of 1b strains from year 2003 to 2014. A second cluster consisting of 4k
211 strains exhibited some degree of substructure, separating samples taken during treat-
212 ment (year 2003) to the other samples. An interpretation of this result was that 1b
213 strains had mainly evolved through drift after treatment, whereas 4k strains might
214 had experienced other evolutionary changes, suggesting selection on this genotype
215 during the evolution of the disease (Caporossi *et al.*, 2019).

216 **Ancient European genomes.** We used PC projections and a Brownian model of
217 factor analysis to study a merged data set consisting of 155k SNP genotypes for 249
218 present-day European individuals and 386 ancient samples from Eurasia. The ages of
219 ancient individuals were less than 12,080 years cal BP, and individuals were selected
220 to be close to present-day Europeans in a preliminary FA analysis to leverage the
221 effect of low genomic coverage on factor one. The data set contained ancient samples
222 mainly from (Olalde *et al.*, 2018; Mathieson *et al.*, 2015; Haak *et al.*, 2015; Mathieson
223 *et al.*, 2018). First, we computed principal components on present-day samples, and
224 projected the ancient samples on the first two PCs (Figure S2). We also computed
225 factors with temporal correction for present-day and ancient samples, choosing the
226 hyper-parameter so that the factors correlate with principal components on present-
227 day individuals ($\lambda = 2 \times 10^{-6}$, Multiple $R^2 = 0.97$ for factor 1, $R^2 = 0.75$ for factor

228 2, $P < 10^{-10}$, Figure S3). Both analyses revealed a similar pattern, in which most
229 ancient samples from Ukraine and all samples from Scandinavia, including hunter-
230 gatherers from Latvia, were close to present-day Finnish samples, ancient samples
231 from Great Britain were close to present-day British samples, and ancient samples
232 from Anatolia and Israel were close to present-day southern Europeans. Ancient
233 samples from Iran, Armenia and Iraq formed a distinct group.

234 Next, we performed an unsupervised time-corrected factor analysis considering
235 ancient samples only. In this analysis, sample ages explained 0.2% of the variance
236 in factor 1 and 5.4% of the variance in factor 2, showing that temporal bias was
237 correctly removed from the first two factors ($\lambda = 10^{-3}$). The FA plot exhibited
238 four main clusters and a pattern of variation strongly consistent with the geographic
239 origin of samples (Figure 5, see Figure S4 for a definition of clusters). A first cluster
240 grouped ancient samples from Ukraine, Latvia and Sweden (Figure 5, green color).
241 Ages in the Scandinavian cluster 1 were around 7,671 years BP (mean value, SD
242 = 1,710 years). A second cluster grouped ancient samples from Russia, including
243 samples from Samara of the Yamnaya culture, Central Europe and Great Britain
244 (Figure 5, dark blue stars to golden points). Ages in cluster 2 were around 3,832
245 years cal BP (SD = 1,570 years). A third cluster grouped individuals from Anatolia
246 and Israel, Southern Europe and Great Britain (Figure 5, brown stars to golden
247 points). Ages in the southeastern cluster 3 were around 6,079 years cal BP (SD =
248 1,311 years). A fourth cluster grouped samples from Central Asia (salmon triangles).
249 Samples from Bronze age Great Britain (4,300 years BP) were grouped in cluster 2,
250 whereas samples from the neolithic period and from the same region were found in
251 cluster 3 representing Southeastern Europe. More generally, samples with ages older

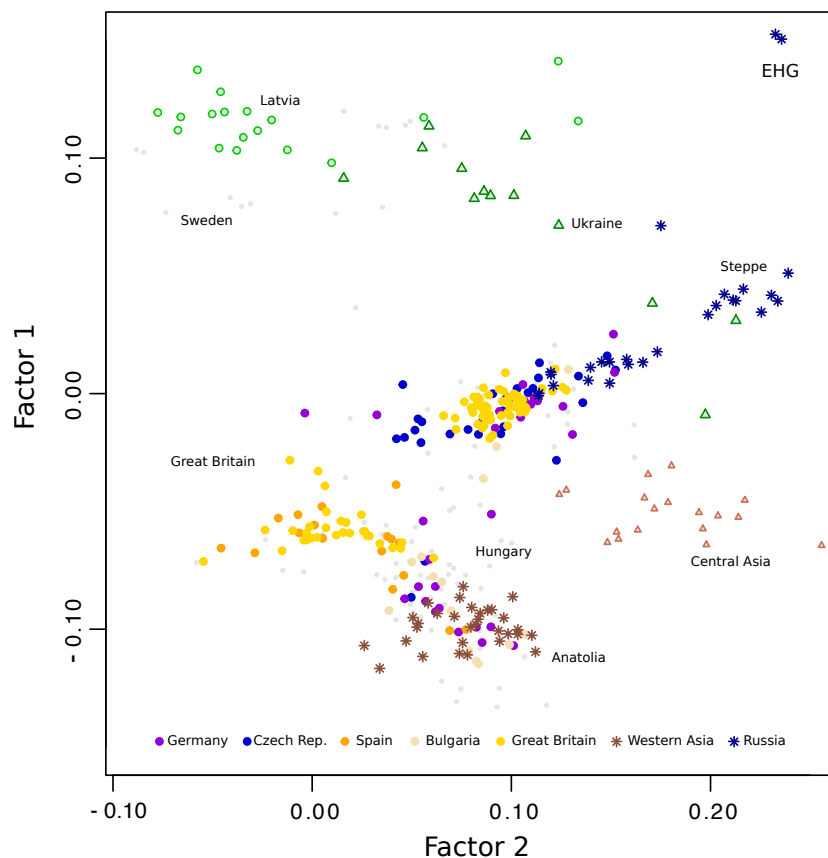


Figure 5. Ancient European genomes. Factor analysis of 386 ancient Eurasian individuals with ages ranging between 400 and 12,000 years BP. Four main groups represent individuals from 1) Northern Europe and Ukraine (green color), 2) Russia, Steppe, Central Europe and the British Isles (average dates around 4k years BP, blue color), 3) Near East, Southern Europe and the British Isles (average dates around 6k years BP, brown color), and 4) Central Asia (Salmon color).

252 than around 4,500 years BP were grouped in the southeastern cluster, while more
253 recent samples from the bronze age clustered with ancient North Eurasian, Russian
254 and Steppe samples in the central cluster 2. Discontinuities in ancestry reflected
255 in factor 1 were observed for samples from Great Britain, Germany and Hungary
256 (Figure 6). In Hungarian samples, a linear trend was observed for the period 4,500 -
257 8,000 years BP, consistent with levels of hunter-gatherer ancestry detected in (Lipson
258 *et al.*, 2017).

259 To assess the genetic ancestry of samples from Great Britain, a second FA was
260 performed. This analysis isolated British samples from ancient North Eurasians
261 and ancient Near Easterners, considered as putative source populations (Figure 7).
262 British samples with dates earlier than 4,300 years cal BP clustered with samples
263 from the Near East. Samples with dates around 4,300 years cal BP (early bronze
264 age) were close to samples from Russia, and a genetic discontinuity was observed with
265 more ancient samples from Anatolia. Estimating admixture coefficients from factor
266 1, the early bronze age samples shared around 64% of their ancestry with the North
267 Eurasian samples and 36% with the Neolithic Easterners. Samples from the middle
268 bronze age (around 3,300 BP) formed a distinct group, suggesting a more complex
269 history than two waves of invasions in the British Isles.

270 Finally, a larger set of 697 ancient samples was considered for replication of PC
271 projection and unsupervised FA results. PCA and FA plots yielded similar descrip-
272 tions of the data when a larger set of ancient samples was considered (Figure S5-S6).

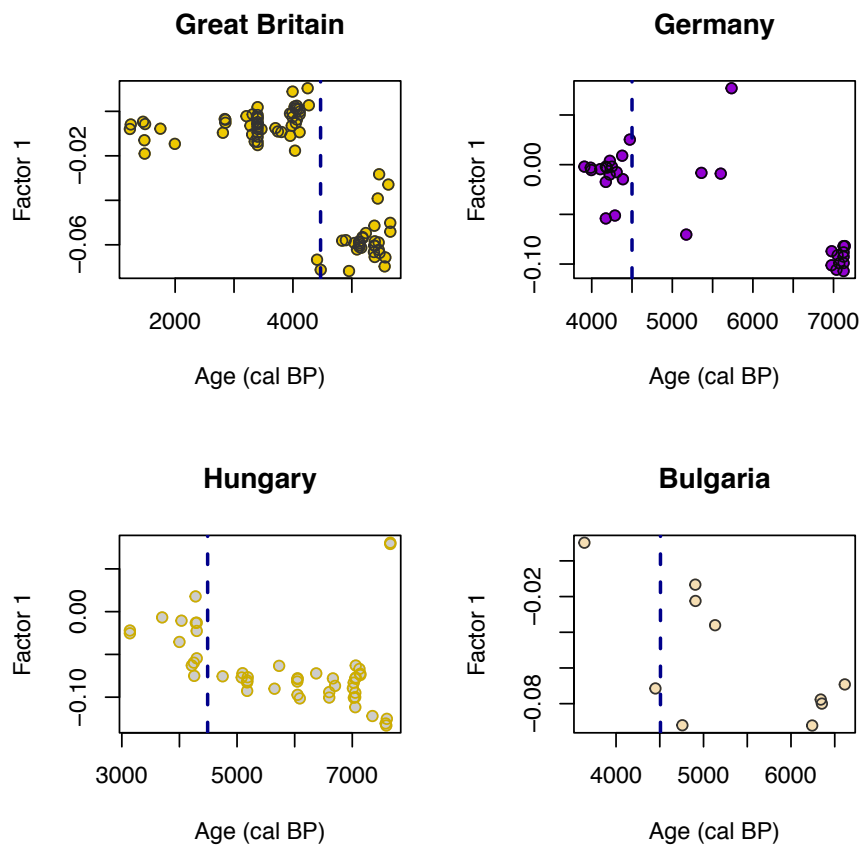


Figure 6. Factor 1 as a function of age (years cal BP). Factor 1 of a temporal FA displayed as a function of age for samples from Great Britain, Germany, Hungary and Bulgaria. The data support a major change in genetic mixture of individuals from Great Britain, Germany, Hungary around 4,500 years BP (dashed line).

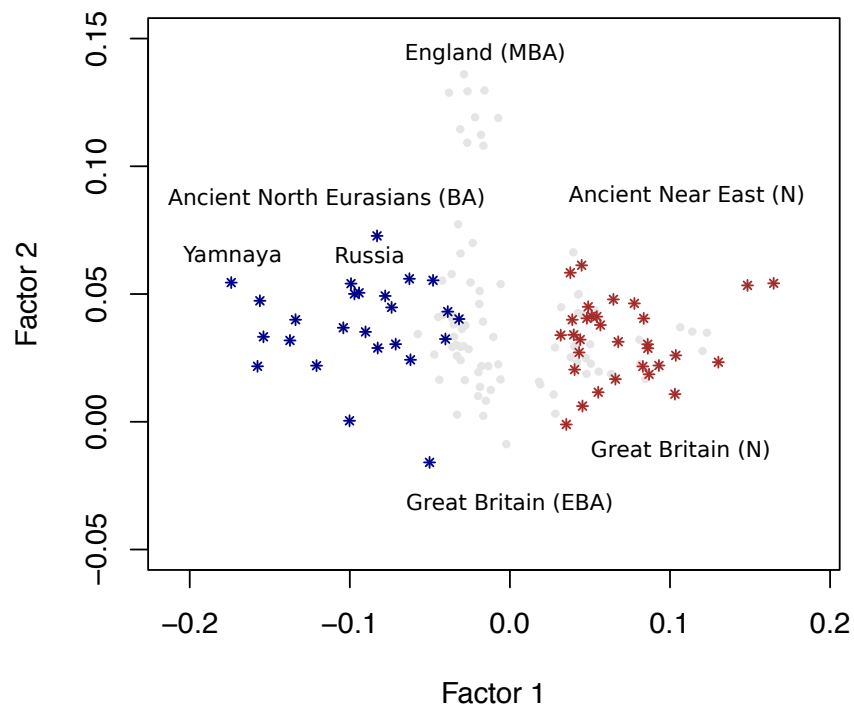


Figure 7. Ancestry of ancient samples from Great Britain. FA plot for samples from ancient North Eurasians (Russia and Samara with Yamnaya Culture, dark blue color), from ancient Near East (Neolithic Anatolia and Israel, brown color), and from Great Britain (grey color). The samples from Great Britain cluster in three groups: Neolithic (N), Early Bronze Age (EBA), Middle Bronze Age (MBA).

273 4 Discussion

274 We introduced a new factor analysis method for describing ancestral relationships
275 among DNA samples collected at distinct time points in the past. Like in PCA, the
276 method is based on a factorial decomposition of the data matrix into a product of
277 score and loading matrices. The most important difference with the PCA approach
278 is that individual scores in FA were corrected for the effect of temporal drift in allele
279 frequency. Based on a diffusion approximation, we approximated allele frequency
280 drift by a Brownian process, and an efficient algorithm based on the singular value
281 decomposition computed the factor estimates.

282 Using a Brownian model of genetic drift, we compared the results of FA with
283 those of PCA and PC projections in simulations of divergence and admixture. In
284 divergence scenarios, distortions due to temporal drift were removed in FA. Correct-
285 ing for temporal drift revealed hidden population structure better than did a PCA.
286 In admixture scenarios, estimates of ancestry coefficients were more accurate in FA
287 than those inferred from principal components. In those simulations, correcting for
288 temporal drift allowed a better representation of admixed individuals than PC pro-
289 jections.

290 Next, we applied temporal corrections to study the evolution of hepatitis C virus
291 in a patient infected by multiple strains. After correction for temporal drift, viral
292 strains clustered according to their phylogenetic classification. In agreement with the
293 fact that the patient did not respond to treatment, FA suggested that 1b strains had
294 mainly evolved through drift after treatment. Evidence for substructure within 4k
295 samples suggested an action for other evolutionary processes among those strains.
296 Caporossi *et al.* (2019) reported that nucleotide diversity was higher in 1b time sam-

297 ples than in 4k time samples, which might indicate that drift was more important in
298 the 4k population. With the FA result, this suggests that distinct corrections should
299 be applied to 4k and 1b samples. We performed a separate FA with 4k samples only
300 (not shown), and the observed substructure persisted. Overall, the FA plot supported
301 the hypothesis that drift was not the only process acting on the genetic diversity of
302 4k genotypes, and that those strains might have experienced some form of selection
303 during the course of disease evolution Caporossi *et al.* (2019).

304 In a re-analysis of a merged data set of ancient DNA filtering out SNPs with high
305 levels of missing data and genomes of low coverage, we implemented correction for
306 temporal drift to describe ancestry in samples from ancient Europeans and Eurasians.
307 After correction, the patterns observed in FA plots were consistent with those ob-
308 served in projections of ancient samples on axes built on the 1,000 Genomes data.
309 The factor analysis supported the hypothesis that a major change in genetic mixture
310 of individuals occurred in Great Britain and in continental populations around 4,300
311 years BP (Olalde *et al.*, 2018). Observed FA patterns were more consistent with
312 geography in than those in PC projections, suggesting a role of localized gene flow
313 unseen in previous analyses at the continental scale. Our analysis provided a visual
314 representation of Bronze age British samples consistent with the proportion of North
315 Eurasian and steppe ancestry of the original (Olalde *et al.*, 2018).

316 In conclusion, including corrections for temporal drift resulted in an algorithm
317 with a computational cost similar to a PCA. Determining the model hyper-parameter
318 was based on simple approaches, computing a correlation between sample dates and
319 first FA scores. Our study showed that the FA method corrected biases observed in
320 PC plots successfully. A useful and important feature of the new approach was to

321 avoid supervised analyses in which unbalanced samples over-representing present-day
322 individuals are utilized. The unsupervised approach based on FA revealed details of
323 population structure masked in PC projections, and was generally more accurate
324 than principal component analysis of population structure for ancient samples.

325 5 Materials and Methods

326 **Coalescent simulations.** We used the computer program *msprime* to simulate
327 temporal samples for individuals at distinct time points in the past (Kelleher *et al.*,
328 2016). Firstly, a single population of $N_e = 10,000$ individuals was simulated during
329 4,000 generations. An individual was sampled every 100 generations, resulting in 41
330 samples with ages ranging between 0 (present-day) and 4,000 generations. A total of
331 around 9,000 SNPs were simulated for each individual. Secondly, a divergence model
332 was considered in which an ancestral population of effective size $N_e = 10,000$ split
333 into two sister populations of equal sizes 1,500 generations ago. Twenty-four individ-
334 uals with ages ranging from 0 to 1000 generations were sampled every 100 generations
335 (four present-day individuals were simulated), and around 8,800 SNPs were simulated
336 for each individual. One hundred replicate data sets were created with the same de-
337 mographic parameters. For each simulation, the Davies-Bouldin index was computed
338 (Davies and Bouldin, 1979). The Davies-Bouldin index is a metric for evaluating the
339 degree of clustering in multidimensional data, and ranges between zero and one. Cor-
340 rections for temporal drift in allele frequency are expected to provide index values
341 closer to one than those for principal components. Thirdly, an admixture model was
342 considered in which an ancestral population of effective size $N_e = 10,000$ split into
343 two sister populations of equal sizes 1,300 generations ago. The two divergent pop-

344 ulations came into contact 500 generations ago, and this event gave rise to a third
345 population. Individuals in the admixed population shared 75% ancestry with the
346 first ancestral population, and 25% ancestry with the second ancestral population.
347 One hundred individuals were sampled from the admixed present-day population,
348 and fifty individuals were sampled from each ancestral population, 1,000 generations
349 ago. A total of around 9,600 SNPs were simulated for each individual. One hundred
350 replicate data sets were created with the same demographic parameters. For each
351 simulation, we computed the centers of the ancestral and admixed population on the
352 first axis, and we estimated admixture proportion based on the ratio of distances
353 between population centers. We also did this for the first factor with correction for
354 temporal drift. We eventually computed mean squared estimation errors both for
355 PCA and for FA estimates.

Generative model simulations. Since the correction method is not restricted
application to ancient DNA, we performed a series of experiments using the generative
model defined in equation (1)

$$\mathbf{Y} = \mathbf{UV}^T + \epsilon, \quad \epsilon \sim \mathbf{N}(0, \alpha^{-1}\mathbf{C} + \sigma^2\mathbf{I}).$$

356 The objective was to evaluate statistical errors for latent factor estimates in a general
357 context. The generative model simulations have the advantage of creating artificial
358 data for which the ground truth is available. Based on a genealogical interpretation
359 of principal components, we devised two series of simulations (McVean, 2009). The
360 first scenario considered a divergence model in which two populations evolved without
361 gene flow. In this case, populations were grouped separately along the first factor, and
362 their divergence time was represented by the distance separating the group means.

363 The samples were taken at random times in the past and correlated noise was included
364 in the data matrix. The second scenario considered an admixture model in which two
365 populations diverged in the distant past and an admixture event occurred recently.
366 Half of the samples were ancient, taken from the ancestral populations at random
367 times in the past, and the other half of the samples were collected from the admixed
368 population in present time.

369 For the divergence model, the factor matrix \mathbf{U} contained $K = 3$ factors, simulated
370 as Gaussian independent random variables. The standard deviation for first factor,
371 s_1 , measured divergence between the two ancestral populations, and was varied in the
372 range from 2 to 10. Factors 2 and 3 had lower standard deviations, respectively
373 equal to $s_2 = 1.5$ and $s_3 = 0.5$, so that \mathbf{u}_1 contained the largest genomic information.
374 The λ parameter, representing an inverse temporal signal-to-noise ratio, was chosen
375 in the range $[10^{-1}, 10^{-6}]$. The number of samples, n , was equal to 200, and the
376 number of markers was kept to $p = 1,000$. Loadings were simulated as independent
377 standard Gaussian random variables, and the residual variance was set to $\sigma^2 = 1$.
378 For each simulation the squared correlation between the true \mathbf{u}_1 and estimated factor
379 $\hat{\mathbf{u}}_1$ was computed.

380 For the admixture model, the factor matrix \mathbf{U} contained $K = 3$ factors. In the
381 first factor, the two ancestral populations were positioned (with a standard deviation
382 of 1) so that the distance separating their centers, d_1 , measuring divergence between
383 them, was varied in the interval $[10,12]$. Factors 2 and 3 had standard deviations equal
384 to $s_2 = 1.2$ and $s_3 = 1$. Admixed individuals were positioned so that center was at
385 relative distance a from ancestral population 1, and $1 - a$ from ancestral population 2,
386 where a represents the ancestry contribution of population 1 to modern samples. The

387 simulated ancestry coefficients ranged between $a = 0.2$ and $a = 0.4$. The λ parameter
388 was set to $\lambda = 5.10^{-2}$. The number of samples was set to $n = 200$, and the number
389 of markers was kept to $p = 5,000$. The loadings were simulated as independent
390 Gaussian, $N(0,0.2)$, random variables, and the residual error was set to $\sigma = 0.1$. We
391 performed a total of 100 simulations. For each simulation, the squared correlation
392 between the true \mathbf{u}_1 and estimated factor $\hat{\mathbf{u}}_1$ was computed, and an estimate of the
393 ancestry coefficient was provided, based on the relative positions of cluster means in
394 $\hat{\mathbf{u}}_1$.

395 **Hepatitis C virus data.** To understand chronic infection in non-responder hep-
396 atitis C virus (HCV) patients treated with dual therapy in the 2000's, Caporossi *et al.*
397 (2019) performed deep sequencing on the NS5B (381 bp) region of the viral genome
398 for a patient followed at Grenoble-Alpes University Hospital. The patient had a
399 known date of infection because of an identified transmission event due to transfu-
400 sion. The patient was treated with dual therapies based on pegylated interferon and
401 ribavirin. The treatment had been administered for six months from January to June
402 2003, and a total of height serum samples were available for a follow-up period of 13
403 years. Co-infection by viral genotypes 4k and 1b was detected, and $n = 1,934$ RNA
404 samples from years 2002 to 2014 were studied.

405 **Ancient Human DNA samples.** A merged data set consisting of genotypes for
406 1,820 ancient and present-day individuals compiled from published papers was down-
407 loaded from David Reich's repository (<https://reich.hms.harvard.edu/>). The
408 downloaded data matrix contained up to 1.23 million positions in the genome. Con-
409 sidering age defined as average of 95.4% date with range in cal BP computed as 1950

410 CE, Eurasian samples with age less than 12,080 years were retained. The data matrix
411 was filtered out for samples falling far outside of the present-day Europeans in a pre-
412 liminary FA analysis, leading to a median genomic coverage of 3.35x and a minimum
413 coverage of 0.51x in the final data set. Only genomic positions with less than 25% of
414 missing genotypes were analyzed. Missing genotypes were imputed by using a matrix
415 completion algorithm based on sparse non-negative matrix factorization (Frichot and
416 François, 2015; Frichot *et al.*, 2014).

417 The resulting data set contained 155,682 genotypes for 249 present-day European
418 individuals from the 1,000 Genomes project (phase 3) and 386 ancient samples from
419 Eurasia studied in previous works (The 1000 Genomes Project Consortium, 2015)
420 (Supplementary File 1). The most important contributions to samples included in
421 our data set were 1) 137 ancient individuals in (Olalde *et al.*, 2018) including 72
422 individuals from Great Britain, 30 from Czech Republic, 24 from Hungary, 14 from
423 Germany and 13 from Russia, 2) 74 ancient individuals in (Mathieson *et al.*, 2015) (31
424 same samples with 390k in (Haak *et al.*, 2015)), including 49 individuals from Great
425 Britain, 15 from Turkey, 35 from Finland, 8 from Russia, 3) 57 ancient individuals
426 from (Mathieson *et al.*, 2018), including 18 individuals from Great Britain, 11 from
427 Hungary, 7 from Germany, 6 from Finland, 11 from Russia, 5 from Ukraine, 4) 40
428 ancient individuals in (Lipson *et al.*, 2017), including 6 individuals from Great Britain,
429 9 from Hungary, 14 from Finland, 4 from Ukraine. For a full list of individuals
430 studied see Table S2. A larger set of genotypes with 5,081 ancient and present-day
431 individuals from the same repository was also considered in analyses. Following the
432 same filtering and imputation procedures as for the first data set, the resulting data
433 contained 123,763 genotypes for 477 present-day European individuals from the 1k

434 Genomes project and 697 ancient samples from previous studies (Supplementary File
435 2). The data were imputed from genotypes with 20% missing SNPs.

436 **Acknowledgements.** We thank the Paris-Saclay Center for Data Science 2.0 (IRS)
437 and LRI for funding BD and SL salaries and supporting FJ’s mobility. We also thank
438 Cyril Furtlehner for discussions. This article was developed in the framework of the
439 Grenoble Alpes Data Institute, supported by the French National Research Agency
440 under the “Investissements d’avenir” program (ANR-15-IDEX-02).

441 **References**

442 Caporossi, A., Kulkarni, O., Blum, M. G. B., Leroy, V., Morand, P., Larrat, S.,
443 François, O. (2019). Using high-throughput sequencing for investigating intra-host
444 hepatitis C evolution over long retrospective periods. *Infection, Genetics and Evo-*
445 *lution*, 67, 136-144.

446 Carroll, M. W., Matthews, D. A., Hiscox, J. A., Elmore, M. J., Pollakis, G., Ram-
447 baut, A., *et al.* (2015). Temporal and spatial analysis of the 2014-2015 Ebola virus
448 outbreak in West Africa. *Nature*, 524(7563), 97.

449 Caye, K., Jumentier, B., Lepeule, J., François, O. (2019). LFMM 2: Fast and accu-
450 rate inference of gene-environment associations in genome-wide studies. *Molecular*
451 *Biology and Evolution* 36 (4), 852-860.

452 Caye, K., Jay, F., Michel, O., François, O. (2018). Fast inference of individual ad-
453 mixture coefficients using geographic data. *The Annals of Applied Statistics*, 12(1),
454 586-608.

- 455 Davies, D. L., Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transac-*
456 *tions on Pattern Analysis and Machine Intelligence*, (2), 224-227.
- 457 Diaconis, P., Goel, S., Holmes, S. (2008). Horseshoes in multidimensional scaling and
458 local kernel methods. *The Annals of Applied Statistics*, 2(3), 777-807.
- 459 Duforet-Frebourg, N., Slatkin, M. (2016). Isolation by distance and time in a
460 stepping-stone model. *Theoretical Population Biology*, 108, 24-35.
- 461 Frichot, E., Schoville, S. D., Bouchard, G., François, O. (2012). Correcting principal
462 component maps for effects of spatial autocorrelation in population genetic data.
463 *Frontiers in Genetics*, 3, 254.
- 464 Frichot, E., Schoville, S. D., Bouchard, G., François, O. (2013). Testing for associa-
465 tions between loci and environmental gradients using latent factor mixed models.
466 *Molecular Biology and Evolution*, 30(7), 1687-1699.
- 467 Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., François, O. (2014). Fast and
468 efficient estimation of individual ancestry coefficients. *Genetics*, 196(4), 973-983.
- 469 Frichot, E., François, O. (2015). LEA: an R package for landscape and ecological
470 association studies. *Methods in Ecology and Evolution*, 6(8), 925-929.
- 471 Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., *et*
472 *al.* (2015). Massive migration from the steppe was a source for Indo-European
473 languages in Europe. *Nature*, 522(7555), 207.
- 474 Halko, N., Martinsson, P. G., Tropp, J. A. (2011). Finding structure with random-

- 475 ness: Probabilistic algorithms for constructing approximate matrix decomposi-
476 tions. *SIAM review*, 53(2), 217-288.
- 477 Harris, A. M., DeGiorgio, M. (2017). Admixture and ancestry inference from ancient
478 and modern samples through measures of population genetic drift. *Human Biology*,
479 89(1), 21-47.
- 480 Hill, M. O., Gauch, H. G. Jr. (1980). Detrended Correspondence Analysis: an im-
481 proved ordination technique. *Vegetatio*, 42,47-58.
- 482 Joseph T.A., Pe'er I. (2018) Inference of population structure from ancient DNA. In:
483 Raphael B. (eds) *Research in Computational Molecular Biology. RECOMB 2018.*
484 Lecture Notes in Computer Science, vol 10812. Springer, Cham, Switzerland.
- 485 Kalaitzis, A., Lawrence, N. (2012). Residual component analysis: Generalising pca for
486 more flexible inference in linear-gaussian models. arXiv preprint arXiv:1206.4560.
- 487 Kelleher, J., Etheridge, A. M., McVean, G. (2016). Efficient coalescent simulation
488 and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5),
489 e1004842.
- 490 Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Prob-*
491 *ability*, 1(2), 177-232.
- 492 Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge Univer-
493 sity Press, Cambridge, UK.
- 494 Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., *et al.*

- 495 (2014). Ancient human genomes suggest three ancestral populations for present-
496 day Europeans. *Nature*, 513(7518), 409.
- 497 Lee, S., Zou, F., Wright, F. A. (2010). Convergence and prediction of principal com-
498 ponent scores in high-dimensional settings. *Annals of Statistics*, 38(6), 3605.
- 499 Lipson, M., Szécsényi-Nagy, A., Mallick, S., Pósa, A., Stégmàr, B., Keerl, V. J., *et*
500 *al.* (2017). Parallel palaeogenomic transects reveal complex genetic history of early
501 European farmers. *Nature*, 551(7680), 368.
- 502 Loève, M. (1948). Fonctions aléatoires du second ordre. In *Processus Stochastiques et*
503 *Mouvement Brownien*, P. Levy (ed.), Gauthier-Villars, Paris, France.
- 504 Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg,
505 S. A., *et al.* (2015). Genome-wide patterns of selection in 230 ancient Eurasians.
506 *Nature*, 528(7583), 499.
- 507 Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N.,
508 Mallick, S., *et al.* (2018). The genomic history of southeastern Europe. *Nature*,
509 555(7695), 197.
- 510 McVean, G. (2009). A genealogical interpretation of principal components analysis.
511 *PLoS Genetics*, 5(10), e1000686.
- 512 Novembre, J., Stephens, M. (2008). Interpreting principal component analyses of
513 spatial population genetic variation. *Nature Genetics*, 40(5), 646.
- 514 Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., *et al.*

- 515 (2018). The Beaker phenomenon and the genomic transformation of northwest
516 Europe. *Nature*, 555(7695), 190.
- 517 Patterson, N., Price, A. L., Reich, D. (2006). Population structure and eigenanalysis.
518 *PLoS Genetics*, 2(12), e190.
- 519 Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., *et al.* (2012).
520 Ancient admixture in human history. *Genetics*, 192(3), 1065-1093.
- 521 Peter, B. M. (2016). Admixture, population structure, and F-statistics. *Genetics*,
522 202(4), 1485-1501.
- 523 Pickrell, J. K., Pritchard, J. K. (2012). Inference of population splits and mixtures
524 from genome-wide allele frequency data. *PLoS Genetics*, 8(11), e1002967.
- 525 Pritchard, J. K., Stephens, M., Donnelly, P. (2000). Inference of population structure
526 using multilocus genotype data. *Genetics*, 155(2), 945-959.
- 527 Slatkin, M. (2016). Statistical methods for analyzing ancient DNA from hominins.
528 *Current Opinion in Genetics and Development*, 41, 72-76.
- 529 Slatkin, M., Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the*
530 *National Academy of Sciences*, 113(23), 6380-6387.
- 531 Skoglund, P., Sjödin, P., Skoglund, T., Lascoux, M., Jakobsson, M. (2014). Investi-
532 gating population history using temporal genetic differentiation. *Molecular Biology*
533 *and Evolution*, 31(9), 2516-2527.
- 534 Skoglund, P., Mathieson, I. (2018). Ancient genomics of modern humans: the first
535 decade. *Annual Review of Genomics and Human Genetics*, 19, 381-404.

536 The 1000 Genomes Project Consortium. (2015). A global reference for human genetic
537 variation. *Nature*, 526, 68-74.

538 Yang, M. A., Harris, K., Slatkin, M. (2014). The projection of a test genome onto a
539 reference population and applications to humans and archaic hominins. *Genetics*,
540 198(4), 1655-1670.

541 Wang, C., Zhan, X., Liang, L., Abecasis, G.R., Lin, X.(2015). Improved Ancestry
542 estimation for both genotyping and sequencing data using projection Procrustes
543 analysis and genotype imputation. *American Journal of Human Genetics*, 96, 926-
544 937.

545 **Supplementary Materials**

546 Supplementary tables and figures for "Inference of Population Genetic Structure
547 from Temporal Samples using Bayesian Factor Analysis" by François et al.

Table S1. R code for temporal factor analysis

```
temporal_fa = function(sample_ages, Y, k = 2, lambda = 1e-3){
  # sample_ages: Ages of samples (year BP/BCE or generations)
  # Y: Matrix of fully imputed genotypes
  # k: Number of factors
  # lambda: Hyper-parameter (range: 1e-1 to 1e-6)

  # conversion of ages as elapsed times between 0 and 1
  Y <- t(scale(t(Y), center = TRUE, scale = FALSE))
  var_Y <- apply(Y, 1, FUN = var)
  range_ages <- max(sample_ages) - min(sample_ages)
  t_n <- 1 - (sample_ages - min(sample_ages))/range_ages
  t_n <- min(var_Y) + (max(var_Y) - min(var_Y)) * t_n

  # Brownian covariance model
  n <- length(t_n)
  C <- matrix(NA, n, n)
  for (i in 1:n){
    for (j in 1:n)
      C[i,j] <- min(t_n[i], t_n[j])}

  # Eigenvectors and eigenvalues
  ec <- eigen(C)
  P_n <- ec$vector
  lambda_n <- ec$values

  # New factors
  D <- diag(sqrt(lambda/(lambda_n + lambda)))
  D_inv <- diag(sqrt((lambda_n + lambda)/lambda))
  sv <- svd(D %*% t(P_n) %*% Y, nu = k)
  U_n <- P_n %*% D_inv %*% sv$u %*% diag(sv$d[1:k])
  W_n <- U_n %*% t(sv$v[,1:k])

  # Returns corrected factors U and latent matrix W
  return(list(u = U_n, w = W_n))
}
```

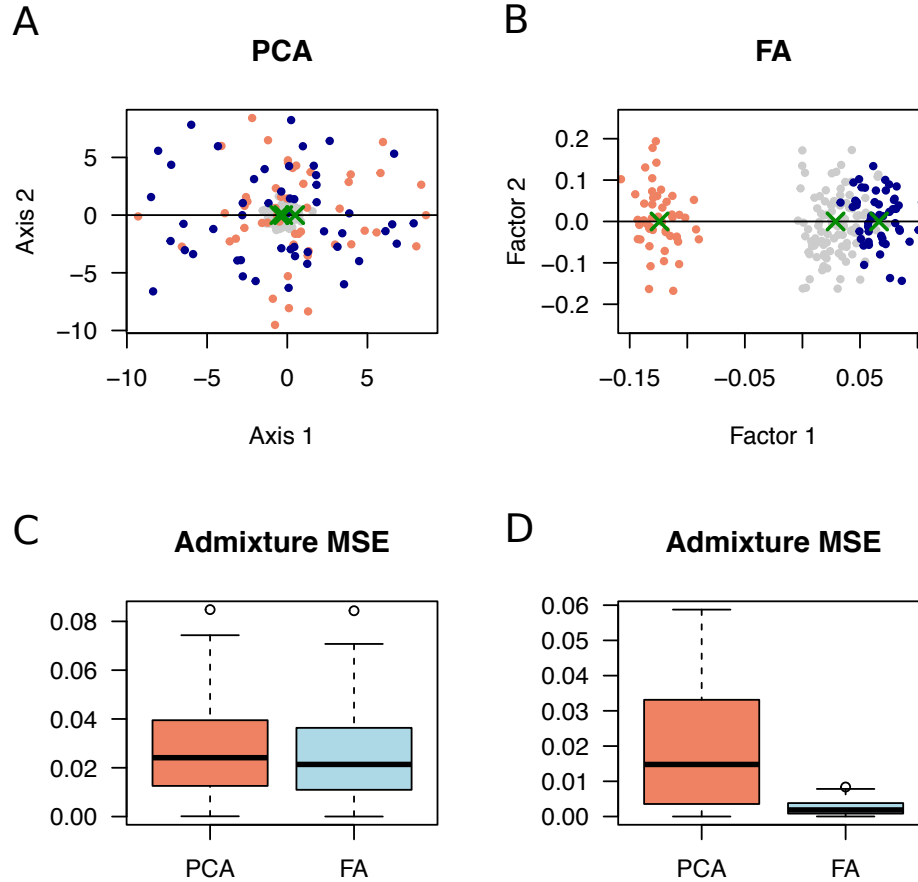
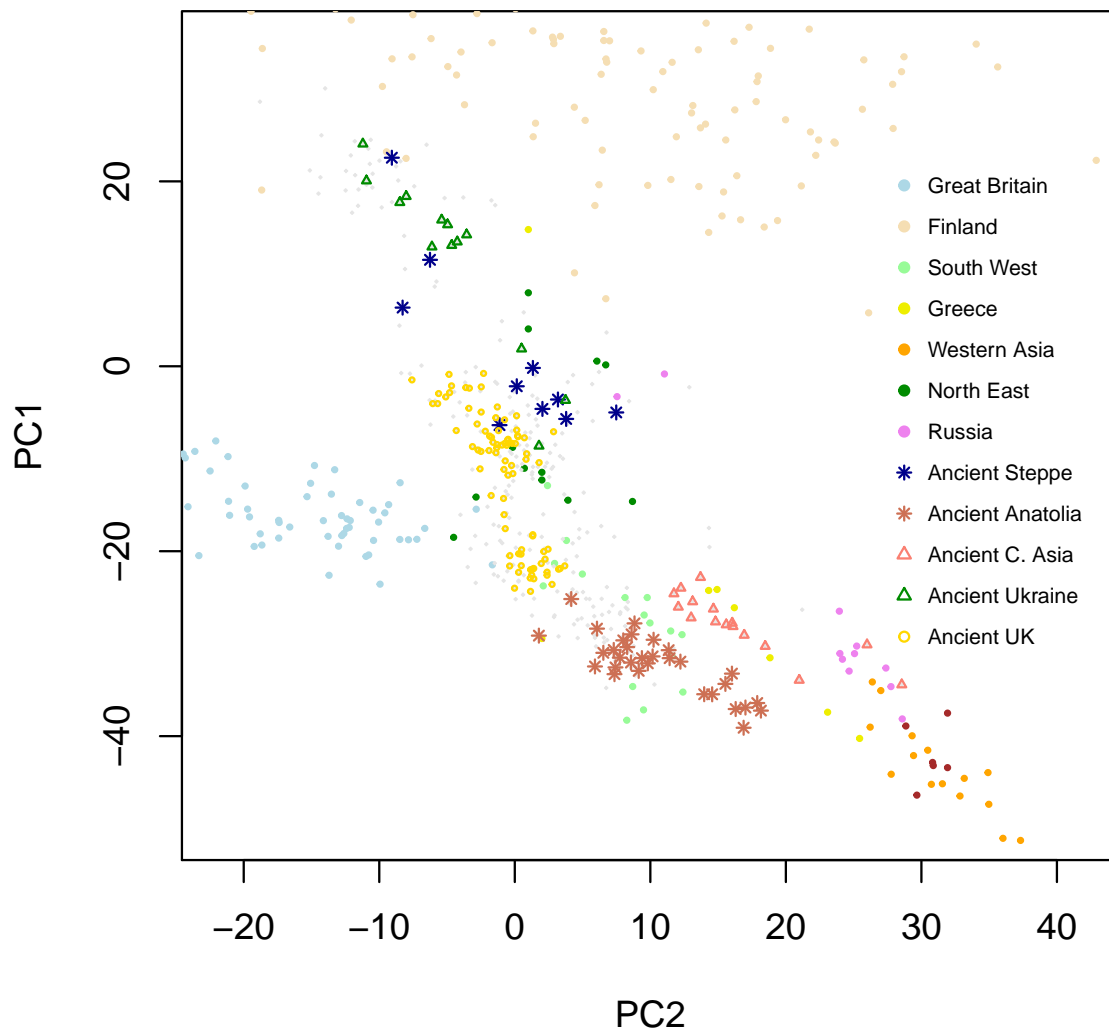


Figure S1. Admixture model simulation. Shrinkage in PC projections. Simulation of two-population admixture models (25-75 % proportions). Two hundred samples with ages equal to 0 (present-day admixed individuals, grey color) and 1,000 generations (ancestors, orange and blue colors) were simulated. A) Plot for PC projection of ancient samples onto the admixed population, with a strong shrinkage effect (coalescent simulation), B) Factor analysis plot showing correction for shrinkage, C) Mean square error for estimates of admixture proportions from PC projections and FA plots (100 generative model simulations), D) Mean square error for estimates of admixture proportions (100 coalescent simulations). Green crosses represent population centers, from which admixture estimates were computed.



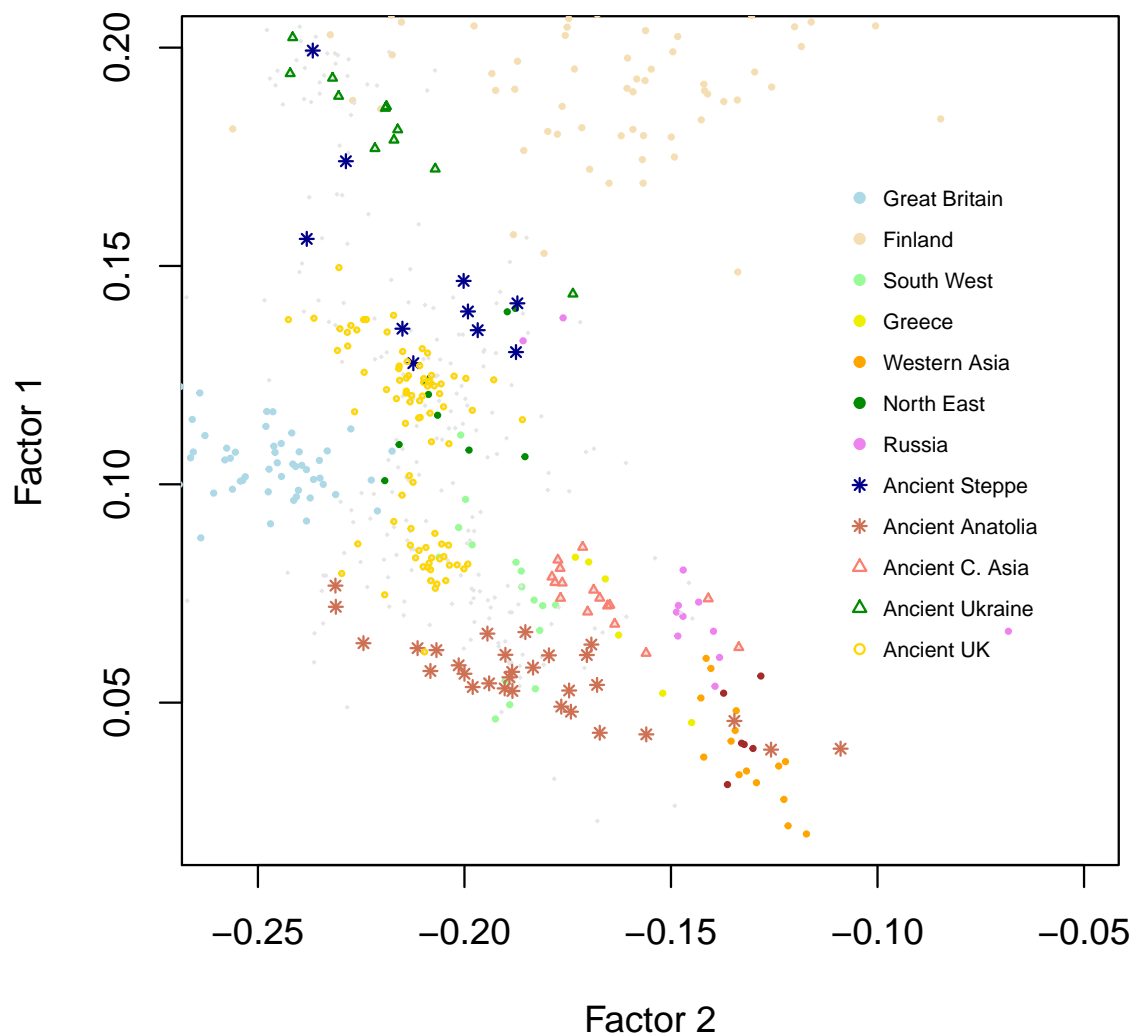


Figure S3. Ancient Humans - Supervised factor analysis. Factor analysis of 386 ancient Eurasian genomes with age ranging between 400 and 12,000 years BP and 249 European genomes from the 1,000 Genomes data. Present-day individuals are represented as colored dots. Smaller light grey dots and other types of dots are ancient genomes.

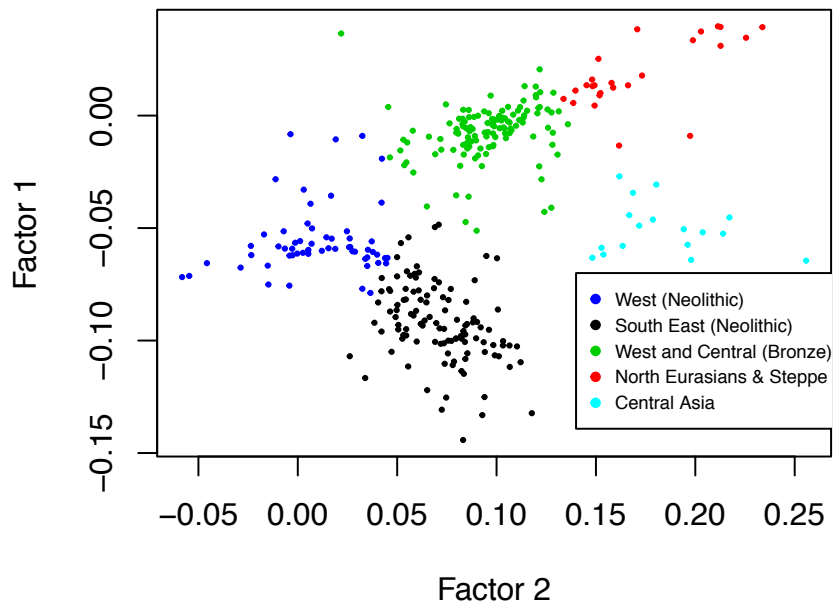


Figure S4. Ancient Europeans - Clusters in factor analysis. Definition of clusters for computing estimates of average ages per factor region. Cluster 1 consists of individuals from Ukraine and Scandinavia, not represented in the plot. Cluster 2 is formed of North Eurasian individuals (Russia, Samara, red color) and central and western Europeans (green color). Cluster 3 is formed of Near Eastern individuals, southern and western Europeans (Neolithic, black and blue colors). Cluster 4 is formed of central Asians (Neolithic, light-blue color). Clustering was performed with a k -means algorithm.

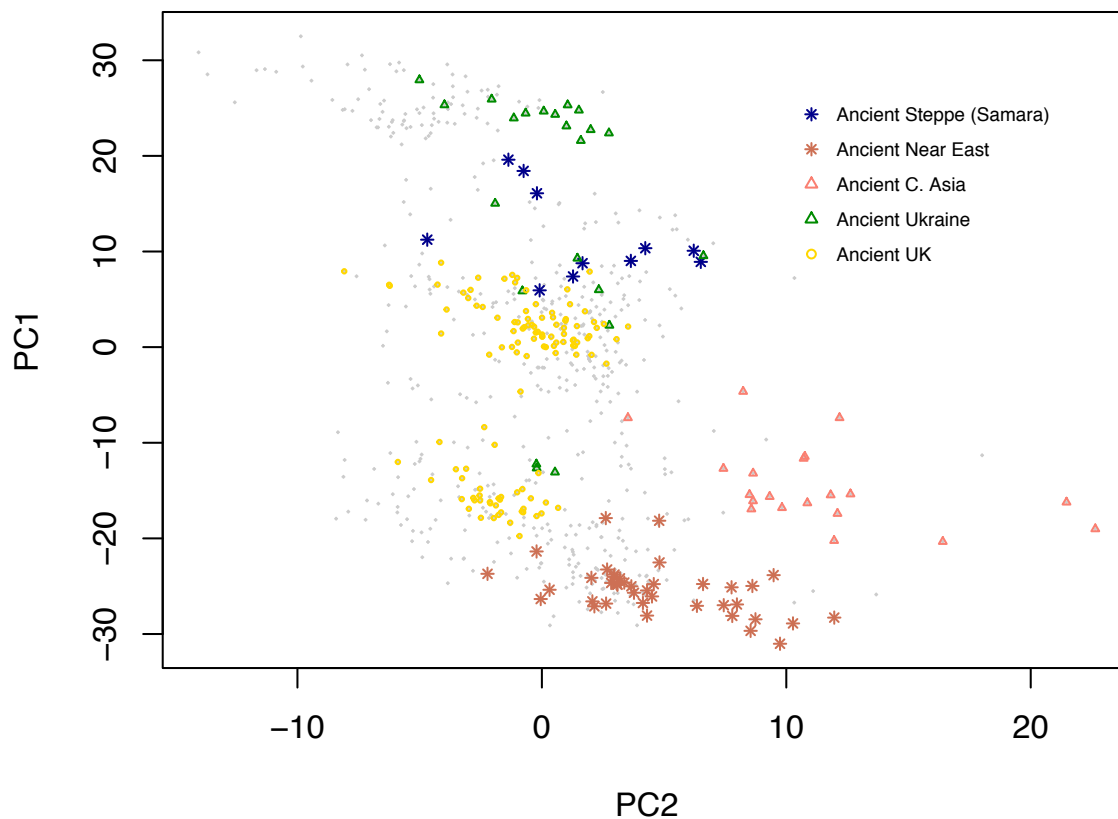


Figure S5. Extended ancient genome data set - PC projections on 1k Genomes data. Projections of 697 ancient genomes on the principal components of 477 genomes from the 1k Genomes data. Only ancient individuals are displayed with some populations emphasized (dates more recent than 12 ky cal BP).

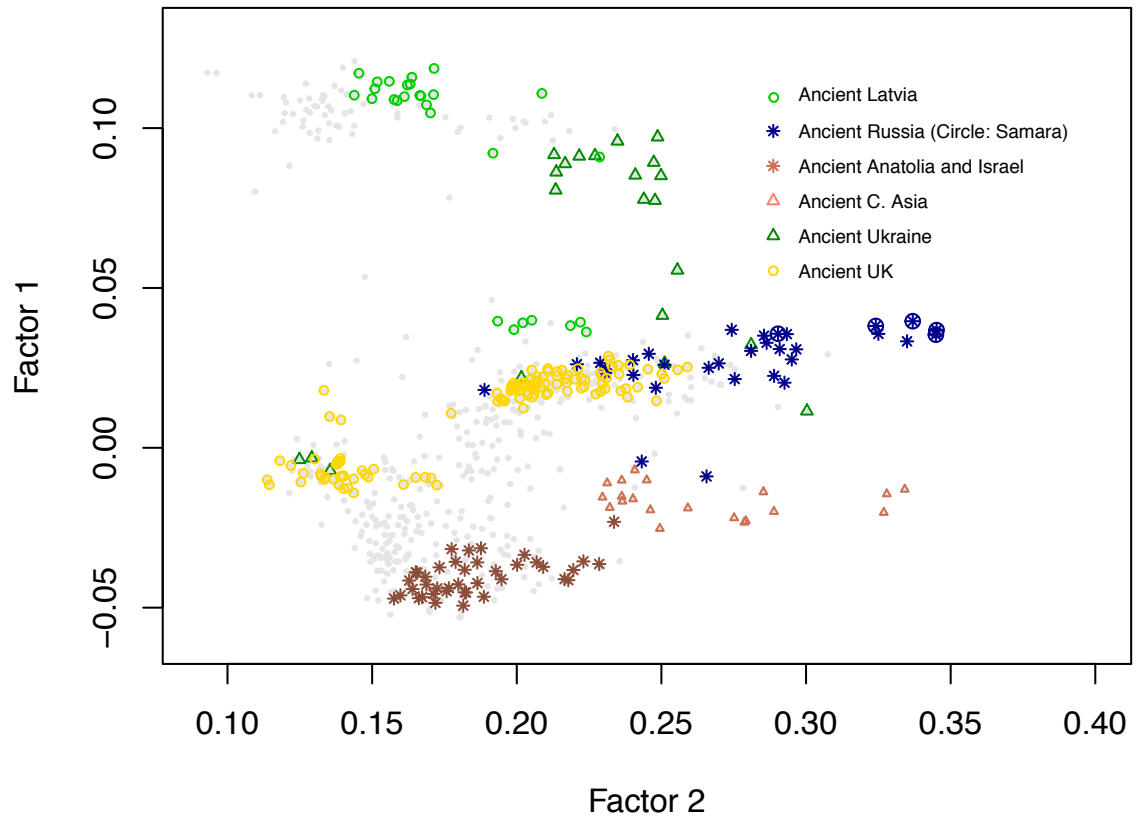


Figure S6. Extended ancient genome data set - Factor Analysis. Factor analysis of 697 ancient genomes with some populations emphasized (dates more recent than 12 ky cal BP). The observed pattern similar to PC projections, but more consistent with geography.