OXFORD

Briefings in Bioinformatics, 2022, 1–16
https://doi.org/10.1093/bib/bbac202
Problem Solving Protocol

KLFDAPC: a supervised machine learning approach for spatial genetic structure analysis

Xinghu Qin, Charleston W. K. Chiang and Oscar E. Gaggiotti

Corresponding authors: Xinghu Qin, Centre for Biological Diversity, Sir Harold Mitchell Building, University of St Andrews, Fife, KY16 9TF, UK & CAS Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences & China National Center for Bioinformation, Beijing, 10010, China. Tel: +0086-13002182629; Fax: 86-10-84097720; Email: qin.xinghu@163.com; Oscar E. Gaggiotti, Centre for Biological Diversity, Sir Harold Mitchell Building, University of St Andrews, Fife, KY16 9TF, UK; Tel: +44 (0)1334 463513; Fax: +44 (0)1334 463513; Email: oeg@st-andrews.ac.uk.

Abstract

Geographic patterns of human genetic variation provide important insights into human evolution and disease. A commonly used tool to detect and describe them is principal component analysis (PCA) or the supervised linear discriminant analysis of principal components (DAPC). However, genetic features produced from both approaches could fail to correctly characterize population structure for complex scenarios involving admixture. In this study, we introduce Kernel Local Fisher Discriminant Analysis of Principal Components (KLFDAPC), a supervised non-linear approach for inferring individual geographic genetic structure that could rectify the limitations of these approaches by preserving the multimodal space of samples. We tested the power of KLFDAPC to infer population structure and to predict individual geographic origin using neural networks. Simulation results showed that KLFDAPC has higher discriminatory power than PCA and DAPC. The application of our method to empirical European and East Asian genome-wide genetic datasets indicated that the first two reduced features of KLFDAPC correctly recapitulated the geography of individuals and significantly improved the accuracy of predicting individual geographic origin when compared to PCA and DAPC. Therefore, KLFDAPC can be useful for geographic ancestry inference, design of genome scans and correction for spatial stratification in GWAS that link genes to adaptation or disease susceptibility.

Keywords: machine learning, population structure, individual geographic origin

Introduction

The genetic differentiation and substructure of human populations are impacted by spatially heterogeneous landscapes [1, 2], social stratification [3, 4], as well as culture [5]. For a long time, an interesting debate in population genetics is whether continuous clines or discrete clusters can better characterize human genetic variation [6–9]. However, without doubts, human population genetic structure exhibits a strong spatial pattern due to population history. On the global scale, this spatial pattern has been described by the isolation-by-distance model, where genetic differentiation between populations increases with increasing geographic distance, as a result of within-population genetic drift and reduced exchange of migrants between populations [10]. Recent studies, mostly at a continental scale (i.e. Europe, Asia), have shown that genetic variation significantly aligns with geography and exhibits spatial patterns that can be inferred by principal component analysis (PCA) [11–13] and model-based analyses [14, 15]. Some studies have reported that the geographical spread of alleles favoured by natural selection contribute to local adaptation [16]. On the other hand, alleles underlying human complex diseases such as cancer, schizophrenia and heart disease also exhibit geographic patterns [17, 18]. Therefore, successful detection of the genetic structure and correct inference of the individual geographic origin will be helpful for applications to personalized medicine, anthropology and forensics.

To date, several non-parametric approaches have been developed to make inferences about the genetic structure of populations and detect loci under selection. PCA is one of the most widely used approaches for these purposes [19]. The link with population structure was demonstrated by a series of studies [11, 20–22], which gradually showed that the proportion of the

Xinghu Qin is a CAS Special Research Associate at CAS Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences & China National Center for Bioinformation. His research mainly focuses on machine learning and deep learning for population genetic inference.
Charleston W. K. Chiang is an Assistant Professor in the Department of Population and Public Health Sciences at Keck School of Medicine and Department of Quantitative and Computational Biology at Dornsife College of Letters, Arts and Sciences, University of Southern California, USA. He is broadly interested in using genetic approaches to understand how natural selection and demographic history shaped the variations in complex traits within and between diverse human populations.
Oscar E. Gaggiotti is a MASTS Professor at the Center for Biological Diversity, School of Biology, University of St Andrews, UK. His research focuses on the study of spatial patterns of genetic diversity to better understand the evolutionary and ecological processes responsible for their origin and maintenance.
Received: January 12, 2022. **Revised:** April 5, 2022. **Accepted:** April 29, 2022

variance explained by the first principal component (PC) computed from the genome-wide single-nucleotide variation (SNP) genotype matrix is highly correlated with the fixation index, F_{ST}. Furthermore, the linear superposition of PC maps has been used to infer the human geographic origin for various present-day and ancient individuals [11, 18, 23, 24]. However, recent studies reported that under some complex scenarios, PCs are not sufficiently informative to represent population structure, as PCs are linear combinations of the variants without consideration of potential non-linear relationship [25, 26]. It has also been suggested that it might be more robust to use non-linear functions of the top PCs, rather than more PCs, to capture non-linear spatial trends [27].

Despite its widespread use in population genetics, the spatial genetic structures represented by the PCs are, to some extent, not discernible between populations because PCs are a summary of the overall variance lumping together between- and within-population variation [28]. In contrast, Fisher linear discriminant analysis (LDA) [29] can maximize between-group variance while simultaneously minimizing within-group variance. In order to take advantage of this property and fulfil the assumption that variables submitted to LDA are perfectly uncorrelated, Jombart *et al*. proposed DAPC (Discriminant Analysis of Principal Components) [28], a hybrid statistical technique for dimensionality reduction that combines LDA and PCA. DAPC is statistically validated for linear inference and has been successfully applied to study population structure [28, 30]. Nevertheless, individual scores in a population determined by LDA may be subject to bias as LDA assumes equal variance for all populations and weighs individuals in a population using the centroid of the genetic components of that population [31]. This property typically merges samples that might be from multiple populations into a single population. As it is the case with LDA, DAPC does not allow for within-group sub-structuring that may arise through migration or non-random mating [32].

Here, we propose a new method to overcome this limitation, Kernel Local Fisher Discriminant Analysis of Principal Components (KLFDAPC), which follows the same principle as DAPC but uses Kernel Local Fisher Discriminant Analysis (KLFDA) [32] instead of LDA. KLFDA is a more general approach for discriminant analysis that allows not only for within-group sub-structuring (multimodality) but also for non-linear associations among samples (individuals) within groups [32, 33]. Therefore, our method combines non-linear and multimodal feature extraction of KLFDA and the dimension reduction of PCA, which helps overcome some of the limitations of PCA and DAPC.

We compared the performance of KLFDAPC for population structure inference and individual geography prediction with those of PCA and DAPC by applying all three methods to both simulated and empirical datasets. The implementation of our method is freely available in the R package *KLFDAPC* at https://xinghuq.github.io/KLFDAPC/.

Materials and methods

KLFDAPC is aimed at overcoming limitations of other popular dimensionality reduction techniques used to infer the genetic structure of populations. These include the presence of hidden genetic structure and non-linear genetic associations between samples. In principle, this could be achieved using KLFDA, an extension of Fisher discriminant analysis that preserves within-class local structure by evaluating the within- and between-class scatter in a local manner and incorporates non-linear associations using the kernel transformation technique [32, 34]. However, KLFDA works well only for small datasets because the kernel transformation faces two key problems, (i) a heavy computational cost and (ii) large diagonals in the kernel matrix if the number of variables is greater than the number of samples ($P >> n$) [35], which is typically the case for genetic data comprising millions of genetic features (loci). As a solution to this problem, and following the example of DAPC [28], we propose to introduce an initial dimensionality reduction step that captures much of the variance present in the original data. This can be achieved using PCA. Thus, the method we propose integrates dimensionality reduction of PCA and the non-linear feature extraction of KLFDA, making the KLFDAPC scalable to genome-wide variation data.

KLFDAPC can be applied not only to genotype matrices but also to many other types of datasets, such as phenotypic traits and species counts. In what follows, we describe the steps required to implement it in general.

KLFDAPC formulation

The first step in the implementation of KLFDAPC is to obtain the PCs. The detailed description of this step is presented in Supplementary Methods available online at http://bib.oxfordjournals.org/. The next step is to conduct the KLFDA analysis using the first P PCs; exploration and guidelines to choosing the value of P is explained below (see section 'Tuning KLFDAPC parameters'). Here, we focus on describing in detail the formulation of KLFDA. As opposed to the PCA step, in KLFDA, we have to take into account the populations where the individuals were sampled in order to now consider the partitioning of genetic variation into its within- and between-population components. Thus, each individual has a population label $y_i \in (1, 2, \ldots, c)$. Some individuals in a population could be recent migrants, and therefore, the population labels for these individuals might not represent their true source population. KLFDA uses a measure of local affinity (see below) that preserves the within population multimodality while maximizing the between population difference. Therefore, individuals

in the same labelled group but actually from different populations can still be embedded and separated appropriately.

The main objective of KLFDA is to estimate the local Fisher transformation matrix, $\mathbf{T}_{LFDA}$, using the within-population scatter matrix $\overline{\mathbf{S}}^{(w)} \in R^n$ and the between-population scatter matrix $\overline{\mathbf{S}}^{(b)} \in R^n$, and then carry out a generalized eigenvalue decomposition. More details about the kernel local Fisher discriminant analysis formulation can be found in Refs. [32, 33]. Below, we briefly recap the main steps for implementing KLFDA, which include (i) computing the kernel matrix M, and affinity matrix A; (ii) defining the local Fisher transformation matrix $\mathbf{T}_{\mathbf{LFDA}}$ in terms of within- and between-population scatter matrices $\overline{\mathbf{S}}^{(w)}$ and $\overline{\mathbf{S}}^{(b)}$ and (iii) computing an analytical form of the transformation matrix by solving a generalized eigenvalue problem.

Computing the kernel matrix

Once the data have been reduced to *P* PCs, KLFDA first transforms the PC scores into a kernel matrix M via non-linear mapping using a kernel function [36]. In this study, we use a Gaussian kernel, also known as radial basis function kernel. Let $\mathbf{x}_i = (x_{ip})$ and $\mathbf{x}_j = (x_{jp})$ be vectors containing the top *P* PCs for individuals *i* and *j*, for *i* and *j* in {1, …, *n*} and *p* in {1, …, *P*}. The elements of the Gaussian kernel matrix M can be defined as,

$$M_{i,j} = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right), \tag{1}$$

with σ determining the width of the Gaussian kernel [37]. This is a parameter that needs to be tuned, a step that is described in the section 'Tuning KLFDAPC parameters'.

The kernel matrix can be viewed as a *n*-dimensional genetic distance matrix between pairs of individuals where each individual has a population label $y_i \in (1, 2, \ldots, c)$. From the kernel distance matrix, we estimated the genetic affinities between individuals and then used them to calculate the within- and between-population weights.

Computing the affinity matrix

Here, the affinity matrix $A_{i,j}$ between individual *i* and individual *j* is computed using the *k*-nearest neighbour search with the local scaling method [38]. Let $\mathbf{m}_i = (M_{ik})$ and $\mathbf{m}_j = (M_{jk})$ be *n*-dimensional vectors of kernel distances between each one of these individuals and all other individuals calculated from Eq. 1. Let NN_i^K be the set of *K*-nearest neighbours of individual *i* under the Euclidean distance, where *K* is the neighbourhood size. If $\mathbf{m}_i \in NN_j^K$ and $\mathbf{m}_j \in NN_i^K$, *i* and *j* are identified as neighbours; otherwise, they are non-neighbours. The elements of the affinity matrix **A** are given by [38]

$$A_{i,j} = \exp\left(-\frac{\|\mathbf{m}_i - \mathbf{m}_j\|^2}{\sigma_i \sigma_j}\right) \tag{2}$$

where σ_i represents the local scaling of the data samples around $\mathbf{m}_i$, which is determined using

$$\sigma_i = \|\mathbf{m}_i - \mathbf{m}_i^K\| \tag{3}$$

where $\mathbf{m}_i^K$ is the vector of kernel distances for the *K*-th nearest neighbour of *i*. $A_{i,j} \in [0, 1]$ with small values indicating individuals have a low genetic affinity (i.e. are genetically far apart), and larger values indicating high affinity (genetically close individuals).

Calculating the local Fisher transformation matrix T_{LFDA}

As described above, $\mathbf{m}_i$ and $\mathbf{m}_j$ are *n*-dimensional vectors of genetic distances for the *i*-th and *j*-th individuals, respectively. Let n_y represent the sample size for population *y* so that $n = \sum_{y=1}^{c} n_y$. Furthermore, let $\overline{A}_{i,j}^{(w)}$ represent the within-population affinity and $\overline{A}_{i,j}^{(b)}$ represent the between-population affinity. Let $\overline{\mathbf{S}}^m$ be the local mixture scatter matrix defined by $\overline{\mathbf{S}}^m = \overline{\mathbf{S}}^{(w)} + \overline{\mathbf{S}}^{(b)}$. The local within-population scatter matrix $\overline{\mathbf{S}}^{(w)}$ and the local between-population scatter matrix $\overline{\mathbf{S}}^{(b)}$ can be obtained as follows.

$$\overline{\mathbf{S}}^{(w)} = \frac{1}{2}\sum\nolimits_{i,j=1}^{n} \overline{A}_{i,j}^{(w)} \left(\mathbf{m}_i - \mathbf{m}_j\right)\left(\mathbf{m}_i - \mathbf{m}_j\right)^T, \tag{4}$$

$$\overline{\mathbf{S}}^{(b)} = \frac{1}{2}\sum\nolimits_{i,j=1}^{n} \overline{A}_{i,j}^{(b)} \left(\mathbf{m}_i - \mathbf{m}_j\right)\left(\mathbf{m}_i - \mathbf{m}_j\right)^T, \tag{5}$$

$$\overline{A}_{i,j}^{(w)} = \left\{ \begin{array}{c} \frac{A_{i,j}}{n_y}, \text{if } y_i = y_j = y \\ 0, \text{if } y_i \neq y_j. \end{array} \right\}, \tag{6}$$

$$\overline{A}_{i,j}^{(b)} = \left\{ \begin{array}{c} A_{i,j}\left(\frac{1}{n} - \frac{1}{n_y}\right), \text{if } y_i = y_j = y \\ \frac{1}{n}, \text{if } y_i \neq y_j. \end{array} \right\}, \tag{7}$$

As opposed to linear discriminant analysis (LDA) in which the within-group scatter and the between-group scatter are obtained using the group centroids and their overall average, here the scatter matrices in $\overline{\mathbf{S}}^{(w)}$ and $\overline{\mathbf{S}}^{(b)}$ are weighted by the affinities. In this case, genetically distant individuals within a population have less influence on $\overline{\mathbf{S}}^{(w)}$ and $\overline{\mathbf{S}}^{(b)}$.

We define the local mixture scatter matrix as $\overline{\mathbf{S}}^{(m)} \equiv \overline{\mathbf{S}}^{(w)} + \overline{\mathbf{S}}^{(b)}$; thus,

$$\overline{\mathbf{S}}^{(m)} = \frac{1}{2}\sum\nolimits_{i,j=1}^{n} \overline{A}_{i,j}^{(m)} \left(\mathbf{m}_i - \mathbf{m}_j\right)\left(\mathbf{m}_i - \mathbf{m}_j\right)^T, \tag{8}$$

$$\overline{A}_{i,j}^{(m)} \equiv \left\{ \begin{array}{c} \frac{A_{i,j}}{n}, \text{if } y_i = y_j \\ \frac{1}{n}, \text{if } y_i \neq y_j. \end{array} \right\}, \tag{9}$$

Therefore,
$\overline{\mathbf{S}}^{(m)} = \frac{1}{2}\sum_{i,j=1}^{n}\overline{A}_{i,j}^{(m)}(\mathbf{m}_i\mathbf{m}_i{}^T + \mathbf{m}_j\mathbf{m}_j{}^T - \mathbf{m}_i\mathbf{m}_j{}^T - \mathbf{m}_j\mathbf{m}_i{}^T)$
$= \sum_{i=1}^{n}(\sum_{j=1}^{n}\overline{A}_{i,j}^{(m)})\mathbf{m}_i\mathbf{m}_i{}^T - \sum_{i,j=1}^{n}\overline{A}_{i,j}^{(m)}\mathbf{m}_i\mathbf{m}_j{}^T$ (10)
Equation (10) can be expressed in matrix form as

$$\overline{\mathbf{S}}^{m} = \mathbf{M}\overline{\mathbf{L}}^{(m)}\mathbf{M}^T, \tag{11}$$

where $\overline{\mathbf{L}}^{(m)} \equiv \overline{\mathbf{Q}}^{(m)} - \overline{\mathbf{A}}^{(m)}$, and $\overline{\mathbf{Q}}^{(m)}$ is the n-dimensional diagonal matrix with the i-th diagonal elements being $Q_{i,i}^{(m)} \equiv \sum_j^n \overline{A}_{i,j}^{(m)}$. Likewise, $\overline{\mathbf{S}}^{(w)}$ can be expressed as $\overline{\mathbf{S}}^{(w)} = \mathbf{M}\overline{\mathbf{L}}^{(w)}\mathbf{M}^T$, where $\overline{\mathbf{L}}^{(w)} \equiv \overline{\mathbf{Q}}^{(w)} - \overline{\mathbf{A}}^{(w)}$, and $\overline{\mathbf{Q}}^{(w)}$ is the n-dimensional diagonal matrix with the i-th diagonal element being $\overline{Q}_{i,i}^{(w)} = \sum_j^n \overline{A}_{i,j}^{(w)}$.

Using $\overline{\mathbf{S}}^{(w)}$ and $\overline{\mathbf{S}}^{(b)}$, the local Fisher transformation matrix $\mathbf{T}_{\mathrm{LFDA}}$ can be defined as [33],

$$\mathbf{T}_{\mathrm{LFDA}} = \underset{T \in R^{d\times m}}{\operatorname{argmax}} \left[tr\left(\left(\mathbf{T}^T\overline{\mathbf{S}}^{(w)}\mathbf{T}\right)^{-1}\mathbf{T}^T\overline{\mathbf{S}}^{(b)}\mathbf{T}\right)\right], \tag{12}$$

T_{LFDA} is the ratio of between-population ($\overline{S}^{(b)}$) and within-population ($\overline{S}^{(w)}$) variances, also known as the F-statistic in LDA, which is used to find the best transformation matrix to maximize Fisher's criterion [29].

Solution of the eigenvalue decomposition problem to obtain T_{LFDA}

Noting that $\mathbf{M}$ is a symmetric matrix, $\mathbf{T}_{\mathrm{LFDA}}$ is obtained by solving [32, 33],

$$\mathbf{M}\overline{\mathbf{L}}^{(b)}\mathbf{M}\varphi = \lambda\mathbf{M}\overline{\mathbf{L}}^{(w)}\mathbf{M}\varphi, \tag{13}$$

where $\overline{\mathbf{L}}^{(b)} = \overline{\mathbf{L}}^{(m)} - \overline{\mathbf{L}}^{(w)}$.

In practice, Eq. (13) cannot be solved because $\overline{\mathbf{L}}^{(w)}$ is always singular. Therefore, Sugiyama [32] proposes regularizing $\mathbf{M}\overline{\mathbf{L}}^{(w)}\mathbf{M}$ and solving instead.

$$\mathbf{M}\overline{\mathbf{L}}^{(b)}\mathbf{M}\varphi = \lambda\left(\mathbf{M}\overline{\mathbf{L}}^{(w)}\mathbf{M} + \varepsilon\mathbf{I}_n\right)\varphi, \tag{14}$$

where I is the identity matrix and ε is a small constant used to regularize the within population distances to provide a more stable matrix.

Using neural networks to evaluate performance of dimensionality reduction methods and to assign individuals to their geographic origin

The three methods we compared are aimed at selecting and combining the input variables into a reduced number of features that capture most of the genetic structure information present in the original dataset. Each method produces a distinct set of reduced features that cannot be directly compared. However, each set contains information about the geographic origin of individuals in the sample, and therefore, they can be used to assign individuals to source populations or geographic coordinates. Therefore, in order to compare the performance of the three methods, we implemented an artificial neuronal network that uses the top two or three features as predictors of each individual geographic or population origin and then calculates the accuracy of each method to assign individuals of known origin. To achieve this goal, we use synthetic data generated through a simulation study covering a wide range of population structure scenarios (see below).

Based on the above-described rationale, we implemented a neural network (Figure 1) and used it as a classifier to assign individuals to populations or as a regression to predict the individuals' geographic origin (latitude and longitude). Neural networks have been well explained in a series of studies [39–45]. A typical single hidden layer neural network consist of an input layer, a hidden layer and an output layer, with nodes in the hidden layers transforming the information between layers using a non-linear activation function (see Supplementary Information). The neural network is optimized based on a loss function that measures the fit of the predicted output to the true value. In the case of classification, we constructed a single-layer neural network with a logistic activation function to assign individuals to populations and used Shannon entropy as the loss function. In the case of regression problems, we used three hidden layers and a logistic activation function with the mean squared error as the loss function. More details about the neural networks and its tuning and fitting are given in Supplementary Methods available online at http://bib.oxfordjournals.org/.

Simulation study

To compare the performance of KLFDAPC over the existing commonly used approaches (PCA, DAPC), we simulated four scenarios that differ in spatial structure: island model, stepping stone model, hierarchical island model and hierarchical stepping stone model using the coalescent-based simulator fastsimcoal2 [46, 47].

For each model, we simulated 16 populations comprising 2000 haploid individuals (equivalent to 1000 diploid individuals). The island model (Supplementary Figure S1A available online at http://bib.oxfordjournals.org/) (including hierarchical island model, Supplementary Figure S1B available online at http://bib.oxfordjournals.org/) and stepping stone model (Supplementary Figure S1C available online at http://bib.oxfordjournals.org/) (including hierarchical stepping stone model, Supplementary Figure S1D available online at http://bib.oxfordjournals.org/) differ in the composition of aggregates (regions) and migration pattern. The hierarchical island model consists of four regions with each region comprising four populations. The hierarchical stepping stone model consists of two regions with each region comprising eight populations.

Under all scenarios, we simulated 44 independent chromosomes with 100 Kb DNA sequences per chromosome with a constant mutation rate of $u = 1 \times 10^{-8}$

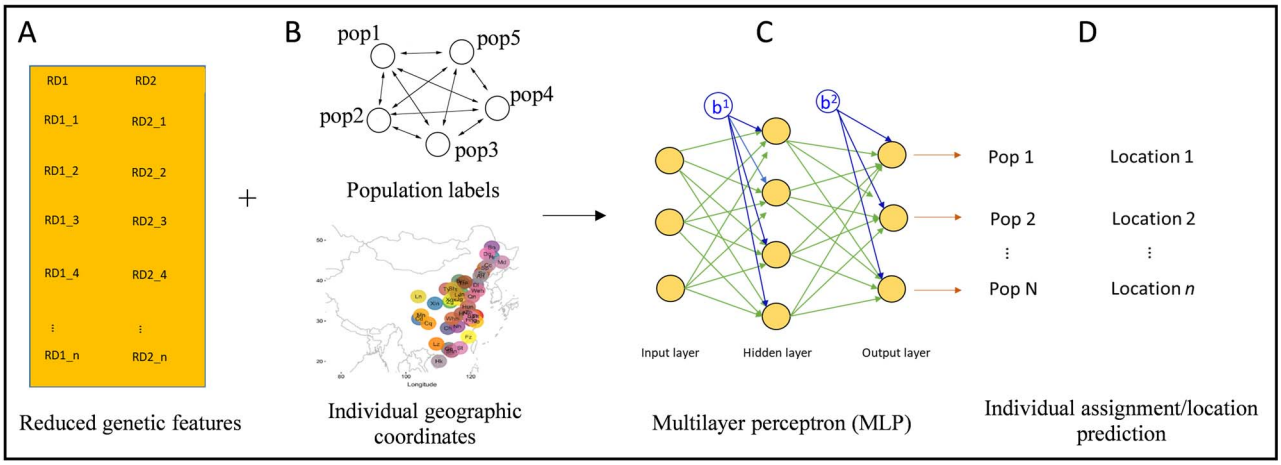

Figure 1. A neural network model for assigning individual membership and predicting the individual geographic coordinates. This framework is based on training a supervised neural network on the reduced genetic features from a dimensionality reduction technique (such as PCA, DAPC and KLFDAPC) given population labels or individual geographic coordinates. The reduced feature matrix ($n \times d$, n is sample size and d is the number of reduced features) obtained from the genetic data are used as the predictor variables (**A**). If the population labels are provided (**B**), they are used as the response variable to carry out classification training through neural network (**C**). The individuals are assigned to the corresponding populations with an optimal neural network model. If the individual geographic coordinates are provided (**B**), the geographic coordinates are used as the response variable to carry out the regression training with neural network (**C**). An optimal neural network model is found and trained to predict the individual geographic coordinates. Finally, the accuracy of the reduced features for assigning individuals to correct populations or for predicting individual geographic coordinates is assessed (**D**) from the optimal neural network model.

per bp per generation, which is typical of humans [48], virus [49], yeast and nematodes [50], and a recombination rate of $r = 1 \times 10^{-8}$ per bp per generation, which is typical of mammals such as humans, but also of plasmids [51], bacteria [52] and human pathogens [53]. In the case of the non-hierarchical scenarios (island model and stepping-stone model), we assumed migration rates between populations to be 0.001, which leads to $Nm = 1$ allowing for the maintenance of polymorphism in each local population thanks to the influx of migrants (c.f. [54]). In the case of the hierarchical island and hierarchical stepping stone models, migration rate between pairs of populations within regions was also 0.001. However, migration between populations from different regions was set to 0.0001 ($Nm = 0.1$) to generate a strong hierarchical subdivision whereby migration from another region cannot, on its own, counteract the effect of genetic drift in any local population. The four scenarios and the respective parameter values used in the simulations are presented in Supplementary Table S1 available online at http://bib.oxfordjournals.org/.

We carried out 10 independent simulations for each scenario and sampled 200 individuals from each population. In total, we obtained 3200 individuals from 16 populations under each spatial scenario. Each scenario generated more than 27 000 polymorphic sites. We removed monomorphic SNPs and filtered the SNPs with a MAF > 0.05 and randomly selected 10 000 sites (biallelic) for downstream analysis.

For each scenario, we first carried out a PCA on the genotype matrix. We then found the number of PC axes and σ that maximized discriminatory power as explained below (see Supplementary Methods available online at http://bib.oxfordjournals.org/). DAPC was implemented using *lda* function from *MASS* package [55], which is initially employed by *dapc* function in *adegenet* package [56]. DAPC and KLFDAPC were conducted using the source population names as the group labels.

Tuning KLFDAPC parameters

Both DAPC and KLFDAPC require to find the optimal number of PCs as input. We chose the number of PCs that has the highest cumulative discriminatory power (accuracy) for population assignment (see above). More precisely, and as described above, we implemented a neural network classifier that in this particular step used PCs as predictive variables to assign individuals to populations under the various spatial scenarios we explored. Then, we evaluated the discriminatory power of the classifier as the number of PCs increased. As Supplementary Figure S2 available online at http://bib.oxfordjournals.org/ shows, the cumulative discriminatory power reached an asymptote as the number of PCs increased from 5 to 20. We found that 20 PCs could correctly discriminate all 16 populations under all scenarios with an accuracy of 1.

Once the optimal number of PCs was found, we then applied a similar procedure to tune σ. In this case, we carried out KLFDAPC analyses of the simulated data based on the 20 top PCs and increasing values of σ (= 0.2, 05, 1, 2, 5). We then used the two and three first reduced features obtained from these analyses as predictive variables of the above described neural network classifier that assigned individuals to populations. When using only two reduced features (Supplementary Figure S3 available online at http://bib.oxfordjournals.org/), the effect of σ is not straightforward. Overall, the power decreases as σ increases but under the island model, it reaches a minimum at $\sigma = 2$ and then increases again, while under the hierarchical island model there is a

steady and rapid decrease in power. On the other hand, under the stepping stone model, the decrease is very slow and the power remains above 90% under all σ values. The decrease in power is less clear under the hierarchical stepping stone model, but here again it remains above 90% under all σ values. Note, however, that when using three reduced features, there is a much clearer pattern of steady decrease in power as σ increases (Supplementary Figure S4 available online at http://bib.oxfordjournals.org/).

Based on this tuning process, we present most KLFDAPC results based on the top 20 PCs and $\sigma = 0.5$ but we also present results for varying σ to visually explore how this parameter affects the clustering of samples in geographic and reduced feature (kernel-induced) space.

Comparing methods performance using simulated data

As mentioned before, we implemented a neural network to assign individuals to source populations based on the top two or three reduced features obtained from each dimensionality reduction method. We then estimated the neural network accuracy and Cohen's Kappa coefficient (κ) to assign individuals to the labelled populations using the information provided by the reduced features (Supplementary Methods available online at http://bib.oxfordjournals.org/).

The use of DAPC and KLFDAPC assumes that population labels assigned to individuals correspond to discrete demographic units. However, in practice, individuals sharing the same habitat patch may represent genetic mixtures, which can introduce errors in the assignment of individuals to populations using the neural network. To compare the performance of DAPC and KLFDAPC under this particularly difficult scenario, we generated synthetic data under a hierarchical island model and then created genetically mixed regions consisting of populations from two different regions (see Figure 4A). The DAPC and KLFDAPC analyses were carried out using the first 20 PCs, and in the case of KLFDAPC, a sigma value of 5.

Application to real datasets

We tested the performance of all three approaches to predict the geographic locations of individuals using two datasets, the European populations from POPRES datasets [57] (dbGaP accession number phs000145.v4.p2) and the Han Chinese populations from CONVERGE data [58] (http://www.ebi.ac.uk/ena/data/view/PRJNA289433). The details on data quality control can be found in Supplementary Methods available online at http://bib.oxfordjournals.org/. To assess the performance of the three approaches (PCA, DAPC and KLFDAPC) in inferring the geographic origin, the predictive performance (R^2 observed values versus the predicted values) between different methods was assessed using model resampling with neural network regression. We also used standard methods to compare the predictive power of PCA, DAPC and KLFDAPC: (i) correlation analysis and (ii) Procrustes analysis. Details of testing using each metric can be found in Supplementary Methods available online at http://bib.oxfordjournals.org/.

Results

Discriminatory power

We assessed the discriminatory power of KLFDAPC for population delineation using the simulated scenarios (Methods and Supplementary Figure S1 available online at http://bib.oxfordjournals.org/), including two non-hierarchical spatial models (the classic island model, Supplementary Figure S1A available online at http://bib.oxfordjournals.org/, and stepping stone model, Supplementary Figure S1C available online at http://bib.oxfordjournals.org/) and two hierarchical spatial models (the hierarchical island model, Supplementary Figure S1B available online at http://bib.oxfordjournals.org/, and hierarchical stepping stone model, Supplementary Figure S1D available online at http://bib.oxfordjournals.org/). We first carried out PCA on the genotype matrix of 3200 sampled individuals under all simulated scenarios. Based on the tuning step described above, we retained the first 20 PCs to conduct the DAPC and KLFDAPC analyses. In addition, we set the KLFDAPC parameter σ to 0.5 (but also present some results using other values). In a first step, we present 2D and 3D plots as the representation of population genetic structure (Figure 2 and Supplementary Figure S5 available online at http://bib.oxfordjournals.org/). We then tested the predictive power of the three approaches to assign individuals to the sampled populations using neural networks (Figure 3).

All methods successfully discriminated between regions under the hierarchical spatial scenarios (hierarchical island model, Figure 2B, F and J, and hierarchical stepping stone model, Figure 2D, H and L). However, PCA and DAPC both failed to clearly delineate local populations under all four scenarios (Figure 2A–H). In contrast, KLFDAPC clearly distinguished genetic stratification among local populations under the stepping stone model based on the first two reduced features (Figure 2K and L) and under the hierarchical stepping stone model based on the first three reduced features (Supplementary Figure S5 available online at http://bib.oxfordjournals.org/). In this latter case, the first reduced feature discriminated between the two higher level regions in the hierarchy while the second and third reduced features together discriminated among populations within regions. The second reduced feature discriminated among populations within region 1 while the third feature discriminated among populations within region 2. Overall, KLFDAPC performed better in identifying the populations under the isolation-by-distance models (the stepping-stone and hierarchical stepping-stone model) than under the island models

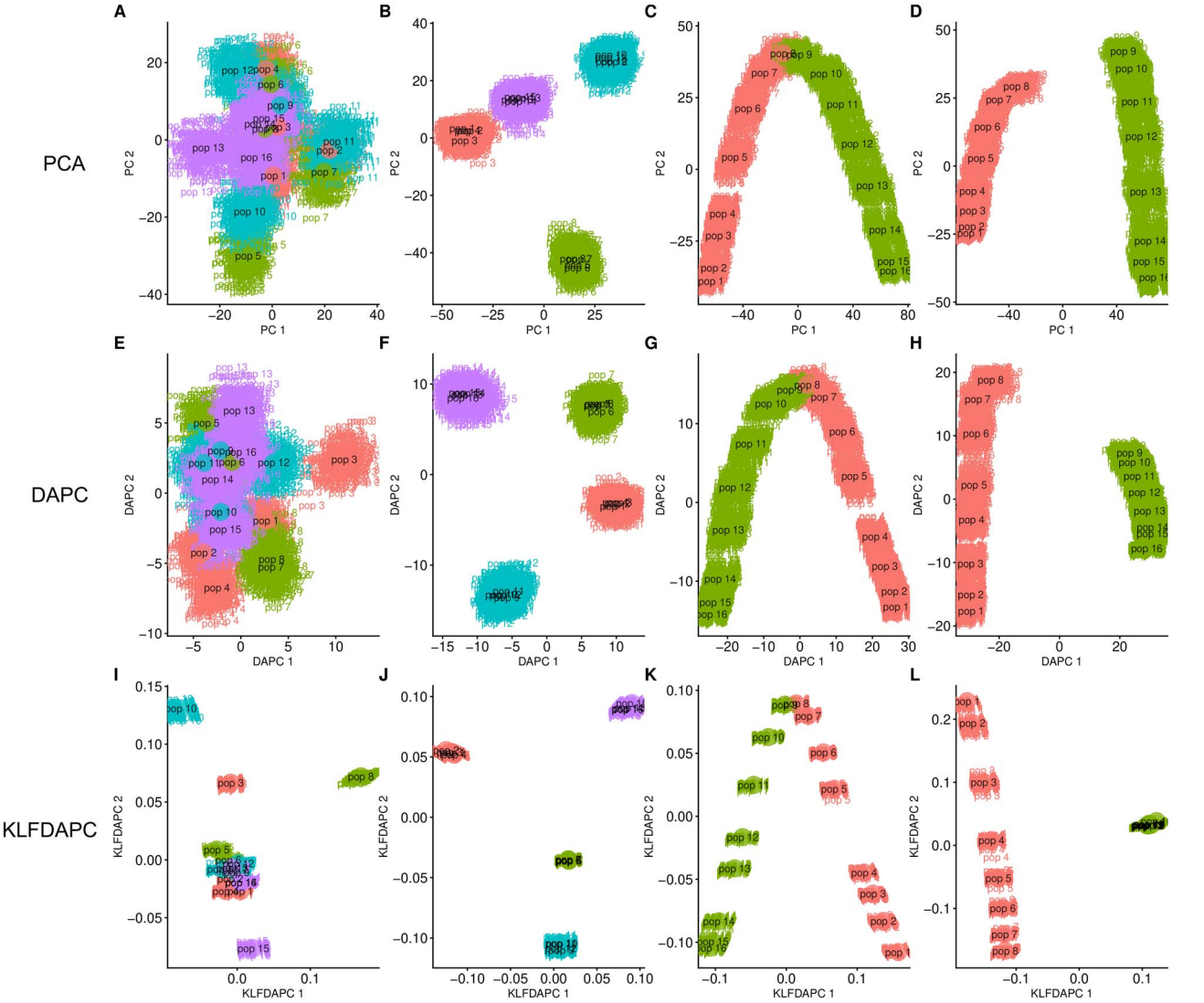

Figure 2. Analyses of simulated data under four spatial scenarios (**A**, **E**, **I**: island model; **B**, **F**, **J**: hierarchical island model; **C**, **G**, **K**: stepping stone model; **D**, **H**, **L**: hierarchical stepping stone model) using PCA, DAPC and KLFDAPC. (**A**–**D**) Genetic structures of four spatial scenarios inferred from PCA; (**E**–**H**) genetic structures of four spatial scenarios inferred from DAPC; (**I**–**L**) genetic structures of four spatial scenarios inferred from KLFDAPC, with $\sigma = 0.5$. The first 20 PCs were used in DAPC and KLFDAPC analyses. The same colour in the scatter plots represents the same region. Individuals are grouped by population names.

(classical island model and hierarchical island model), where populations within regions tend to overlap (Figure 2).

To quantitatively compare the performance of each method (PCA, DAPC and KLFDAPC) in describing genetic structuring, we implemented an artificial neural network that used the first three reduced features obtained from each method to assign individuals to populations (see Methods and Supplementary Information). We then assessed the classification accuracy of each set of reduced features in terms of classification accuracy and Cohen's Kappa coefficient [59]. Consistent with the graphical representation of the spatial structures, the discriminatory accuracy and Cohen's Kappa coefficient (κ) for KLFDAPC were much higher than those achieved by PCA and DAPC under all scenarios (Figure 3). Note that the accuracy and κ of all methods with three reduced features improved over those obtained when only two axes were used (Supplementary Figure S6 available online at http://bib.oxfordjournals.org/) but the strongest improvement was observed for KLFDAPC. We therefore recommend considering more features (i.e. the first three features in Supplementary Figure S5 available online at http://bib.oxfordjournals.org/) to fully characterize population structure under complex scenarios.

As Supplementary Figure S7 illustrates, the pattern of local genetic aggregation is sensitive to the parameter σ. KLFDAPC introduces non-linear genotypic associations using a Gaussian kernel in which σ controls the strength of dispersal/aggregation of the local structure. Lowering the σ values of KLFDAPC could increase the discrimination of discrete clusters, thus increasing the ability of KLFDAPC to delineate distinct aggregates (Supplementary Figs S4 and S7 available online at http://bib.oxfordjournals.org/).

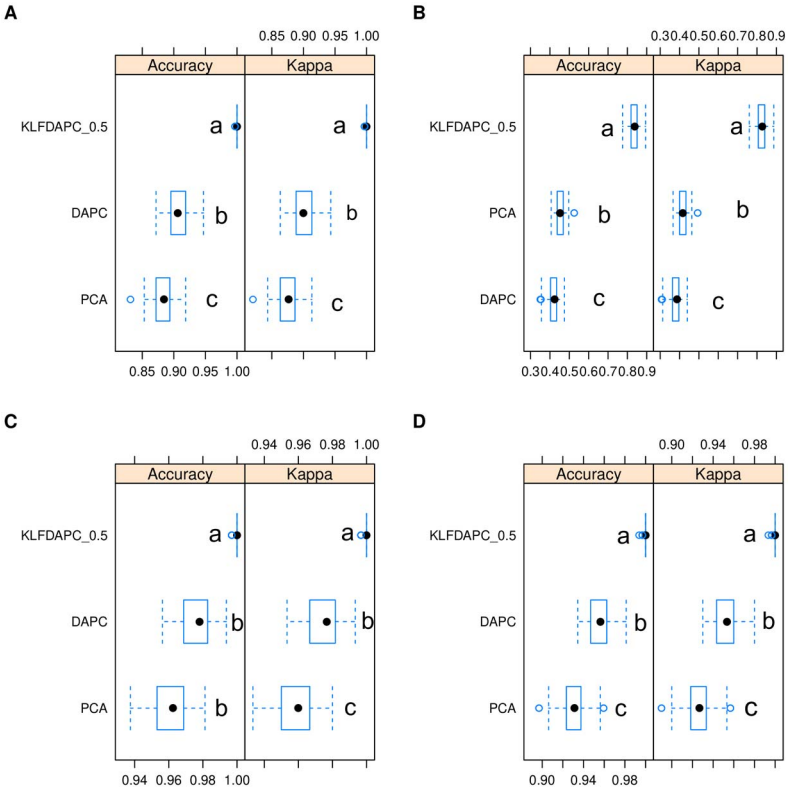

Figure 3. Discriminatory power of three approaches using the first three reduced features as the explanatory variables to distinguish populations. (**A**) Island model, (**B**) hierarchical island model, (**C**) stepping stone model and (**D**) hierarchical stepping stone model. Accuracy and Kappa were estimated after '10-fold-10-repeats' adaptive cross-validation. Comparison between models was tested using a pairwise *t*-test based on results of 100 cross-validation resamples. Different letters indicate the statistical significance at the 0.05 level. *P*-value adjustment: Bonferroni.

In summary, KLFDAPC outperformed PCA and DAPC in discriminating population genetic structure. Furthermore, in the case of stepping stone models, KLFDAPC was able to characterize a spectrum of genetic structure from continuous genetic gradients to discrete clusters with appropriate kernel parameter values (i.e. σ in Gaussian kernel). Therefore, we recommend users to vary kernel parameter values to explore how it influences the results.

The effects of hidden substructure on the delineation of regions under hierarchically structured scenarios

An important problem when using supervised learning is the effect of group mislabelling due to hidden substructure. This can happen, for example when sampling takes place in wintering or feeding areas that can receive migrants from several different regions. To investigate this issue, we considered a scenario where four breeding grounds contributed each to two different feeding grounds (see Figure 4A). Thus, each feeding ground consisted of genetic mixtures of two distinct genetic clusters. Figure 4B shows that DAPC was unable to group individuals according to the region where they bred. On the other hand, KLFDAPC correctly grouped together individuals from the same breeding region. This difference is due to the different way in which the two methods calculate the within-class scatter matrix (i.e. the distance between the position of each sample in multidimensional space and the average position of the class). More precisely, DAPC simply uses the class centroid while KLFDAPC takes into account the genetic affinity between samples.

Analysis of POPRES data

We tested the performance of PCA, DAPC and KLFDAPC to predict the geographic locations of European individuals using POPRES data. The first two PCs of the POPRES data only accounted for 0.29% and 0.15% of the total SNP variance, respectively. These two PCs were remarkably aligned with the map of Europe (Figure 5A and D; Table 1, PC1 versus longitude, $R = 0.872$, PC2 versus latitude, $R = 0.873$), as previously reported [11]. DAPC and KLFDAPC analyses using the top 20 PCs also provided accurate geographic representations of the genetic samples (Figure 5B, C, E and F). Moreover, DAPC and KLFDAPC improved the alignment of genetic samples with their locations. For example, DAPC and KLFDAPC rectified the projected geographic locations between Turkey (TR) and Albania (AL) by bringing them close to each other, while also locating PL (Poland) to its correct position and IR and UK samples closer to their correct location (Figure 5; Supplementary Table S2 available online at http://bib.oxfordjournals.org/, PC1 versus longitude, $R = 0.872$; KLFDAPC1 versus longitude, $R = 0.886$; PC2 versus latitude, $R = 0.873$; KLFDAPC2 versus latitude, $R = 0.934$; the difference in R^2 between PCA and KLFDAPC was significant as indicated in Supplementary Table S2

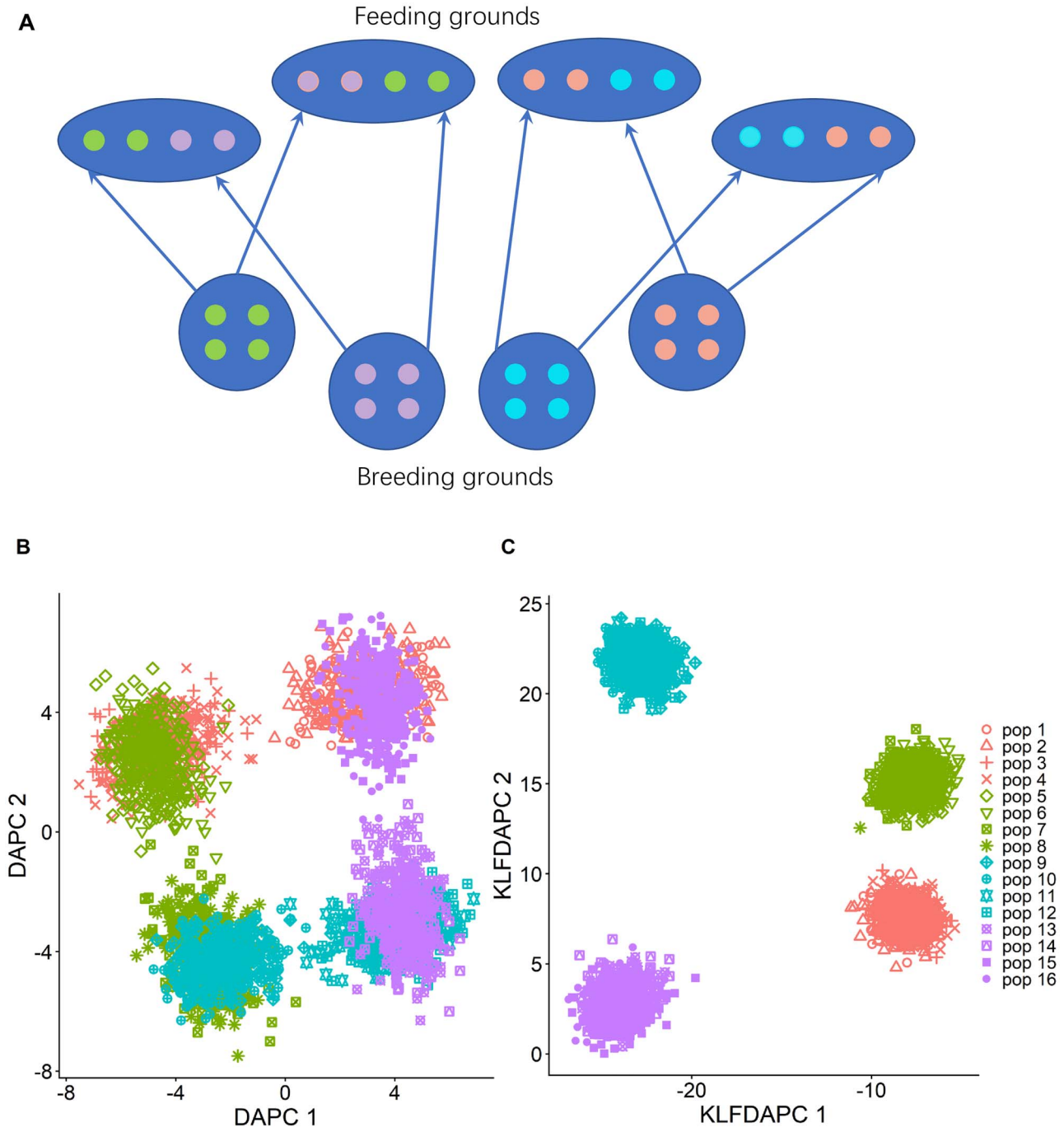

Figure 4. Population structure inference when sampled regions are genetic mixtures. (**A**) Graphical representation where each blue circle represents a region consisting of four breeding grounds. Each blue oval represents a feeding ground composed of individuals from two different regions. Small circles represent populations and are coloured according to the region they belong to. (**B**) Results obtained with DAPC; (**C**) results obtained with KLFDAPC.

available online at http://bib.oxfordjournals.org/). KLFDAPC uses a Gaussian kernel to take into account non-linear associations between samples, where σ determines the decay in the association [32]. Increasing σ from 1 to 5 induced a gradual change of the projected locations between Cyprus (CY) and south-east European countries, as well as Russia (RU) and north-east European countries (Supplementary Figure S9 available online at http://bib.oxfordjournals.org/).

KLFDAPC performed better at predicting the geographic locations of POPRES individuals than PCA and DAPC (Tables 1 and 2; Supplementary Table S2 available online at http://bib.oxfordjournals.org/). Overall, lower σ values tend to aggregate individuals into compact clusters, while high σ values make the individuals more scattered.

When predicting the individual origin via a neural network model using the first two reduced features as the predictor variables, KLFDAPC performed the best among the three methods to predict the individual longitude and the latitude (Tables 1 and 2). DAPC and PCA showed similar power in predicting the individual geographic locations (Tables 1 and 2). Overall, KLFDAPC showed

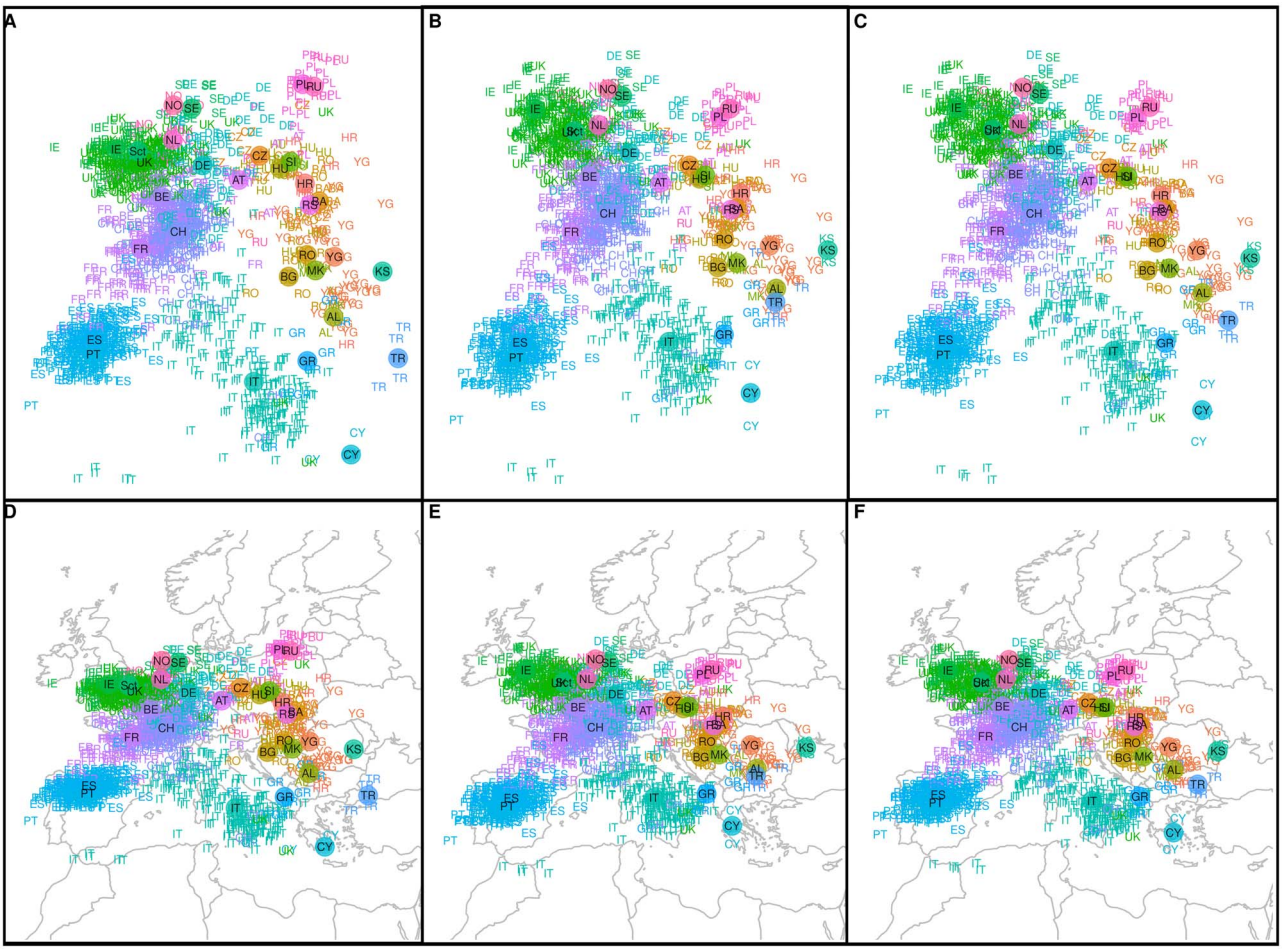

Figure 5. Genetic structure of POPRES dataset represented by the first two reduced features from PCA (**A**), DAPC (**B**) and KLFDAPC (**C**), and projected individual geographic locations within Europe based on PCA (**D**), DAPC (**E**) and KLFDAPC, with $\sigma = 5$ (**F**). The solid circles are the centroid of individuals from the same country. Country abbreviations: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; ES, Spain; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia; Sct, Scotland; SE, Sweden; TR, Turkey; YG, Yugoslavia.

Table 1. Performance of different methods to predict individual locations as measured by a MLP approach with cross-validation (POPRES data)

	Longitude			Latitude			Correlation between observed and predicted location using the optimal MLP model			
Methods	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RD1 versus longitude**	**P value**	**RD2 versus Latitude**	**P value**
PCA	6.683	0.550	5.276	2.283	0.867	1.644	0.756	2.20E−16	0.937	2.20E−16
DAPC	6.956	0.553	5.477	2.313	0.867	1.687	0.542	2.20E−16	0.930	2.20E−16
KLFDAPC (σ =1)	7.333	0.594	5.783	2.794	0.839	2.195	**0.819**	2.20E−16	0.831	2.20E−16
KLFDAPC (σ =2.5)	6.996	0.584	5.430	2.314	0.870	1.706	0.672	2.20E−16	0.873	2.20E−16
KLFDAPC (σ =5)	**6.253**	**0.616**	**4.799**	**2.027**	**0.886**	**1.428**	0.796	2.20E−16	**0.947**	2.20E−16

The best statistic is marked in bold. Abbreviations, RMSE: Root Mean Square Error; MAE: Mean Absolute Error; R^2: R^2 is coefficient of determination.

superior predictive power than PCA and DAPC in predicting individual geographic locations from European populations (Tables 1 and 2).

Traditional summary statistics measuring the performance of the three approaches in inferring individual geographic locations are presented in Supplementary Tables S2 available online at http://bib.oxfordjournals.org/. Procrustes correlation is strongest for KLFDAPC with $\sigma = 2.5$, but the individual correlation between reduced features after Procrustes transformation and latitude or longitude are strongest for KLFDAPC with $\sigma = 1$ (Supplementary Table S2 available online at http://bib.oxfordjournals.org/).

Analysis of CONVERGE data

We also assessed the efficacy of our method to infer the individual geographic origin for a large Han Chinese population from the CONVERGE dataset. The

Table 2. The difference of R^2 (observed value versus predicted value) between different methods estimated by a MLP model for predicting the individual longitude and latitude (POPRES data)

	PCA	DAPC	KLFDAPC_ σ _1	KLFDAPC_ σ _2.5	KLFDAPC_ σ _5
RD1 versus longitude					
PCA		3.47E−06	2.83E−02	−3.14E−03	−1.91E−02
DAPC	1		2.83E−02	−3.15E−03	−1.91E−02
KLFDAPC_ σ _1	**0.0097**	**0.021**		−3.15E−02	−4.61E−02
KLFDAPC_ σ _2.5	1	1	**<0.05**		−1.60E−02
KLFDAPC_ σ _5	**0.011**	**0.026**	**2.26E−05**	0.084	
RD2 versus latitude					
PCA		−0.0038	−0.044	−0.0342	−0.0663
DAPC	1		−0.041	−0.0305	−0.0625
KLFDAPC_ σ _1	0.625	0.570		0.0102	−0.0218
KLFDAPC_ σ _2.5	1	1	1		−0.032
KLFDAPC_ σ _5	**0.01901**	**0.01102**	1	0.41398	

p values marked with bold indicate statistically significant differences between two methods at the $p < 0.05$ level. Upper diagonal: estimates of the difference. Lower diagonal: *P*-value for H0: difference = 0. R^2 (true value versus predicted value) was estimated after 5-fold cross-validations, repeated five times. Comparison between models was tested using a pairwise *t*-test between 100 resamples. *P*-value adjustment: Bonferroni.

CONVERGE data consist of individuals from 24 out of 33 administrative divisions (19 provinces, 4 municipalities and 1 autonomous region) across China. We used the individual-level birthplace information at the province level to denote the geographic origin of each sample [13].

PCA performed poorly in recapitulating the geography of individuals, as PC1 corresponded poorly to both latitude and longitude (Supplementary Table S3 available online at http://bib.oxfordjournals.org/, Figure 6A and D; PC1 versus longitude $R = 0.0508$, PC1 versus latitude $R = -0.351$). However, we can still observe a significant North–South gradient (Supplementary Table S3 available online at http://bib.oxfordjournals.org/, Figure 6A and D; PC2 versus latitude $R = 0.6404$). Compared to PCA, the reduced features obtained from DAPC better represented the genetic gradients along latitude and significantly aligned with East China on the map, where most of the participants were born (Figure 6B and E; Supplementary Table S3 available online at http://bib.oxfordjournals. org/). Notably, KLFDAPC ($\sigma = 0.5$) presented a clear 'boomerang' shape for the genetic structure with Shanghai (Sh), Zhejiang (ZJ), Jiangsu (JS) at the vertex of the boomerang structure. Compared with PCA and DAPC, KLFDAPC with $\sigma = 0.5$ displayed clear correlation with latitude (South–North axis) and aligned significantly better with geography (KLFDAPC2 versus latitude $R = 0.7357$, Supplementary Table S3 available online at http://bib.oxfordjournals.org/, Figure 6 and Supplementary Figure S10 available online at http://bib.oxfordjournals.org/). The spread of samples on the map increases as σ increases, but this made the inferred individual locations inaccurate, for example, as σ increased, individuals were placed out of their birth places (on the sea, Supplementary Figure S10 available online at http://bib.oxfordjournals.org/). Increasing σ values also made the populations indiscernible, especially for populations that are both genetically and geographically related, such as Shanghai, Jiangsu and Zhejiang (Supplementary Figure S10 available online at http://bib.oxfordjournals.org/).

Due to the limitation of sample collection (samples are mainly collected from eastern, coastal China provinces), there was a poor correlation between the first reduced genetic feature and longitude. We note that PC did worse at predicting longitude while DAPC and KLFDAPC improved predictive accuracy but still performed poorly at predicting longitude (Table 3). All methods failed to accurately predict individual longitude using neural networks (Table 3). However, they still performed well in predicting the individuals' latitude (Table 3). The predictive power analysed using neural networks showed that PCA did worst in predicting the individual latitude among these three approaches ($R^2 = 0.727$, Table 3). Even though DAPC ($R^2 = 0.740$) did better than PCA ($P = 5.97E{-}07$), KLFDAPC ($\sigma = 0.5$, $R^2 = 0.767$) performed significantly better than both PCA ($P = 2.2E{-}16$) and DAPC ($P = 2.2E{-}16$) (Tables 3 and 4). In addition, the predictive power R^2 obtained from neural networks is higher than the conventional linear correlation coefficient (Table 3; Supplementary Table S3 available online at http://bib.oxfordjournals.org/).

In summary, KLFDAPC with a σ value of 0.5 outperforms PCA and DAPC in all aspects when predicting the individual geographic origins of Han Chinese people using the CONVERGE dataset (Tables 3 and 4; Supplementary Table S3 available online at http://bib.oxfordjournals.org/), suggesting that genetic features produced by KLFDAPC seem to be a better surrogate for geographic coordinates than PCs.

Discussion

The availability of large genomic databases has pushed researchers to put aside model-based methods in favour of non-parametric approaches, such as PCA [22, 60] and DAPC [28]. In this study, we introduced KLFDAPC, a non-linear approach for inferring population genetic structure and individual geographic origin. Using a neural network with KLFDAPC reduced features as predictive variables, we tested the performance of

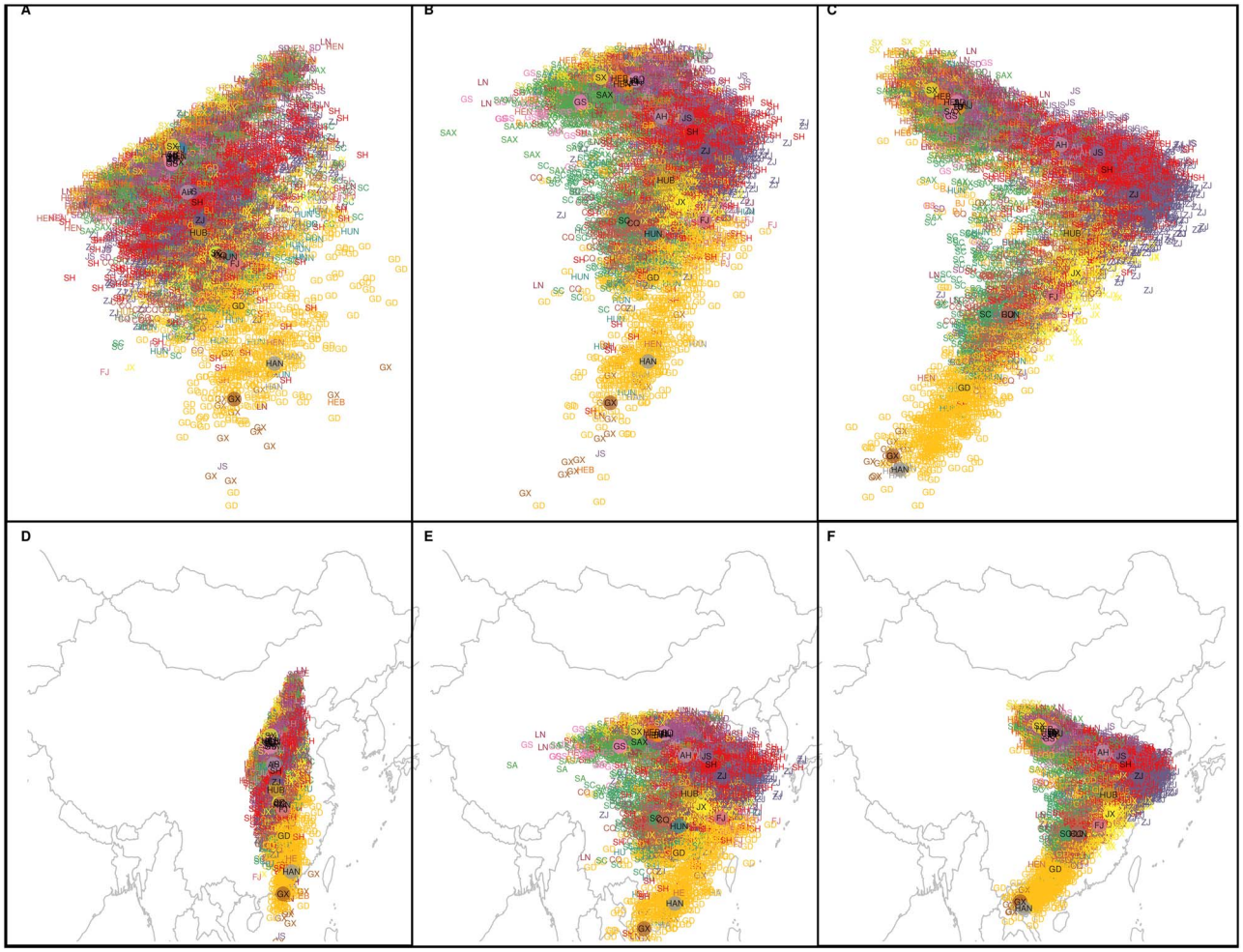

Figure 6. Genetic structure of Han Chinese people from the CONVERGE dataset represented by the first two reduced features from PCA (**A**), DAPC (**B**) and KLFDAPC (**C**), and projected individual geographic locations within China based on PCA (**D**), DAPC (**E**) and KLFDAPC, with $\sigma = 0.5$ (**F**). The solid circles represent the centroid of individuals from the same province. Province abbreviations: Shanghai, SH; Liaoning, LN; Zhejiang, ZJ; Tianjin, TJ; Hunan, HUN; Sichuan, SC; Shaanxi, SAX; Heilongjiang, HLJ; Jiangsu, JS; Shandong, SD; Henan, HEN; Hebei, HEB; Beijing, BJ; Guangdong, GD; Jiangxi, JX; Shanxi, SX; Hubei, HUB; Guangxi Zhuangzu, GX; Chongqing, CQ; Fujian, FJ; Gansu, GS; Jilin, JL; Anhui, AH; Hainan, HAN.

Table 3. Performance of different methods to predict individual locations as measured by a MLP approach with cross-validation (CONVERGE data)

	Longitude			Latitude			Correlation between observed and predicted location using the optimal MLP model			
	RMSE	R^2	MAE	RMSE	R^2	MAE	RD1 versus longitude	*P* value	RD2 versus latitude	*P* value
PCA	37.637	0.003	20.310	5.688	0.530	4.460	NA	NA	0.7274	2.2E−16
DAPC	35.000	0.038	16.641	5.682	0.549	4.460	NA	NA	0.7409	2.2E−16
KLFDAPC_σ_0.5	36.236	0.018	18.071	**5.615**	**0.592**	**4.398**	NA	NA	**0.7670**	2.2E−16
KLFDAPC_σ_1	35.416	0.012	17.403	5.638	0.558	4.442	NA	NA	0.7471	2.2E−16
KLFDAPC_σ_2.5	34.220	0.006	16.625	5.650	0.557	4.457	NA	NA	0.7459	2.2E−16
KLFDAPC_σ_5	34.730	0.015	16.854	5.714	0.553	4.496	NA	NA	0.7427	2.2E−16

The best statistic is marked in bold. Note that models fitted with longitude have larger RMSE and smaller R^2 compared to those fitted with latitude. NA was produced due to the uneven distribution of the samples along longitude when sampled from the space by neural networks. MAE: Mean Absolute Error; RMSE: Root Mean Square Error; R^2: R2 is coefficient of determination.

KLFDAPC for inference of individual population membership and geographic origin. We showed that KLFDAPC outperformed both PCA and DAPC in population structure discrimination and in predicting individual geographic origin using simulated scenarios and empirical population datasets (Figs 2 and 3). Analyses of the POPRES dataset showed that all three methods retrieved a strong correspondence between genetic structure and geography, but KLFDAPC outperformed both other methods in inferring the individual geographic

Table 4. The difference of R^2 (observed values versus predicted values) between different methods estimated by a MLP model for predicting the individual latitude (CONVERGE Data)

	PCA	DAPC	KLFDAPC_σ_0.5	KLFDAPC_σ_1	KLFDAPC_σ_2.5	KLFDAPC_σ_5
PCA		−0.019	−0.062	−0.028	−0.027	−0.023
DAPC	**5.97E−07**		−0.043	−0.009	−0.009	−0.004
KLFDAPC_σ_0.5	**2.2E−16**	**2.2E−16**		0.034	0.035	0.040
KLFDAPC_σ_1	**5.99E−13**	0.450	**2.2E−16**		0.0008	0.0056
KLFDAPC_σ_2.5	**1.12E−12**	0.514	**2.2E−16**	1		0.0048
KLFDAPC_σ_5	**8.93E−13**	1	**2.2E−16**	1	1	

p values marked with bold indicate statistically significant differences between two methods at the $p < 0.05$ level. Upper diagonal: estimates of the difference. Lower diagonal: *P*-value for H0: difference = 0. R^2 (observed values versus predicted values) was estimated after 5-fold cross-validations, repeated five times. Comparison between models was tested using a pairwise *t*-test between 100 resamples. *P*-value adjustment: *Bonferroni*. All models failed to predict the longitude because of the small variance between individuals and poor correlation between features and longitude.

locations (Tables 1 and 2; Supplementary Table S2 available online at http://bib.oxfordjournals.org/). When applying the three methods to the Han Chinese population, PCA exhibited poor performance in characterizing the individual geography. Both DAPC and KLFDAPC provided a much better alignment between genetic structure and geography and remarkably improved the predictive accuracy of individual geographic origin compared to PCA (Figure 6). Again, KLFADPC outperformed both PCA and DAPC in predicting the individual geography in the CONVERGE dataset (Tables 3 and 4). Overall, our study highlights the remarkable performance of KLFDAPC in identifying population genetic structure and in predicting individual geographic origin. We thus propose that KLFDAPC, which extracts the non-linear genetic features and also allows for within-population hidden structuring, may be a useful alternative to PCA and DAPC for many population genetics studies.

There are model-based alternatives to the machine learning methods we considered in our study. However, they generally require users to define a non-linear function modelling the slope of allele frequency across geographic locations [14, 15]. Any pre-specified parametric function is unlikely to sufficiently capture complex geographic patterns in genetic variation, such as multiple modes or peaks in the allele frequency surface as reported by Yang *et al.* [14]. In contrast, KLFDAPC is a non-parametric method that can incorporate both non-linear genetic associations between individuals and hidden sub-structuring within populations. It does not require a non-linear function being defined but only needs to choose an appropriate kernel.

One particular feature of KLFDAPC that differentiates it from PCA and DAPC is the need to specify the value of the kernel parameter, for example σ in the Gaussian kernel used in our study. Our simulation study suggests that values between 0.2 and 5 provide optimal results under most scenarios. However, we consider that it is useful to vary this parameter to explore genetic structuring at different spatial scales. Our simulation study showed that low values of σ help to clearly delineate different regions in hierarchically structured scenarios, while larger values tend to highlight within-region structuring. Also, real-data analyses showed that increasing σ tends to produce more scattered or continuous patterns (Supplementary Figs S9 and S10 available online at http://bib.oxfordjournals.org/). Thus, by tuning σ, it is possible to maximize the power to predict individual geographic origin and, therefore, to better account for population stratification and spatial effects in GWAS analyses.

We found that, in analyses of the CONVERGE data, PCA tended to provide a poor indication of the correct geographic origin of individuals. A previous study of the population structure of Han Chinese people based on the CONVERGE dataset filtered ∼30% of individuals on the basis of poorly imputed genotype in order to reveal strong correlation between the first two PCs and longitude and latitude [13]. In our study, we were able to demonstrate much stronger correlation between reduced genetic features and geography using KLFDAPC compared to PCA, while retaining all 10 461 individuals in the analysis. This result suggests that KLFDAPC is more robust to varying data quality issues such as missing or poorly imputed genotypes than conventional PCA. In particular, missing genotypes is a common problem in genomic studies. The *k*-nearest neighbour algorithm is often used for imputing missing genotypes [61–63]. Note that KLFDAPC uses the *k*-nearest neighbour algorithm (KNN) to calculate the local genetic affinity, which could strongly decrease the influence of genotyping errors on inference. Therefore, KLFDAPC overcomes this artefact experienced by PCA by preserving the non-linear genetic features and local genetic affinity.

A common problem with supervised or population-based methods such as LDA and F_{ST} is the requirement for pre-grouping. Pre-grouping might be subjective or arbitrary, as we rarely know if some individuals in a group might have immigrated recently from other groups or some individuals within a group have unknown origin, which introduce bias to the inference [64] and could mask important evolutionary processes such as migration and cross-breeding [65]. Our results showed that DAPC suffers from this problem as it minimizes the within-group variation based on the group means and can thus lead to erroneous assignment of individual to populations (Figure 4). Thus, extensive genetic mixing or admixing can represent a great challenge for DAPC and lead to inaccurate representations of genetic structure.

Unlike DAPC and other within-group minimizing/averaging approaches, KLFDAPC is less affected by pre-grouping (Figure 4). The main reason is that KLFDAPC computes the within- and between-population affinity based on the KNN weight, minimizing the influence of 'local outliers' (i.e. migrants) on the inference of population structure. Therefore, KLFDAPC can preserve the multimodal genetic structure within populations and overcomes the problems caused by the group-mean (or group-centroid)-based method.

Existing studies propose to use neural network approaches for population assignment [66, 67] and prediction of geographic origin [68] using genetic data. Guinand et al. [66] show that assignment tests based on a neural network classifier with one hidden layer generally outperforms the other methods. In Supplementary Methods available online at http://bib.oxfordjournals.org/, we show that a KLFDAPC-based single-layer neural network largely outperforms the method of Guinand *et al*. [66]. The most likely explanation is that when using individuals' genotypes as input, a single hidden layer may not be capable of extracting all the information present in the raw data. Using instead the reduced features of KLFDAPC overcomes this limitation. Another way of improving the accuracy of individual assignment based on genotype data is proposed by Battey *et al*. [68], who used a neural network with several hidden layers to predict the individual locations based on allele counts and known locations.

Besides predicting the individual geographic origin, we also implemented a genome scan approach to identify genomic regions involved in local adaptation based on KLFDAPC (https://xinghuq.github.io/KLFDAPC/articles/Genome_scan_KLFDAPC.html). Many ecological, evolutionary and medical datasets are complex and may exhibit hidden genetic structuring. We expect that KLFDAPC would be very helpful in these particular situations. First, population structure integrating multimodal structure within large populations could correct spurious results in inferring individual geographic origin (c.f. Han Chinese example in our study). The genetic structure with a non-linear outlier identification model, such as neural network (as opposed to linear regression models in *pcadapt* [69] and LFMM [70]), would provide a better characterization of genomic regions under natural selection or involved in adaptation. Therefore, KLFDAPC could improve the power of detecting loci under selection or involved in local adaptation. Also, KLFDAPC could also be used to correct for stratification in genome-wide association studies, which so far have done so using PCA.

Machine learning algorithms, from simple general linear regression [71], PCA [22, 60], to random forest [72], extreme gradient boosting [73], as well as neural networks [74], have enabled us to capture the systematic signatures of biological or genetic patterns from genomic samples, allowing for the association of genes to phenotypes/diseases and facilitating molecular-based medical applications [75–77]. KLFDAPC represents a new addition to the population genomics toolbox but it is also potentially applicable to other Omics data throughout the biological sciences, including applications in medicine and agriculture.

Key Points

- Most species' genetic diversity exhibits geographic patterns and several methods have been proposed to characterize it. Through a simulation study and real-data analyses, the limitations of the commonly used methods, PCA and DAPC, for spatial genetic data analysis were revealed.
- A supervised machine learning method, KLFDAPC, is introduced to rectify the limitations of PCA and DAPC by capturing non-linear information and preserving the multimodal space of samples.
- KLFDAPC outperformed PCA and DAPC in discriminatory power and in predicting the geographic origin of individuals.
- KLFDAPC can be useful for geographic ancestry inference, design of genome scans and correction for spatial stratification in GWAS.
- KLFDAPC is freely available at https://xinghuq.github.io/KLFDAPC/index.html.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Author contribution

X.Q. and O.E.G. designed the study. C.K.W.C. reviewed the design, acquired the datasets and provided computational resources. X.Q. compiled the package, carried out the analyses and interpreted results with input from C.W.K.C. and O.E.G. X.Q. and O.E.G. wrote the manuscript with the input from C.K.W.C. All authors contributed to editing and revisions of the manuscript.

Funding

CSC-University of St Andrews Joint Scholarship (to X.Q.); International Postdoctoral Exchange Fellowship Program (Talent-Introduction Program) from China Postdoc Council (to X.Q.); National Institute of General Medical Sciences (NIGMS) of the National Institute of Health (grant R35GM142783 to C.W.K.C.). Part of the computation for this work is supported by USC's Center for Advanced Research Computing (https://carc.usc.edu).

Data and code availability

The input files and scripts for generating simulation are available at https://github.com/xinghuq/KLFDAPC/tree/sm/Simulation_files. The POPRES data with request approval #90291-1 is accessible via dbGaP Study accession number phs000145.v4.p2. The CONVERGE data

are available at the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/data/view/PRJNA289433). Scripts for analysing the POPRES and CONVERGE datasets are available at https://github.com/xinghuq/KLFDAPC/tree/sm/Scripts. The package KLFDAPC used for analysis is available at https://xinghuq.github.io/KLFDAPC/.

Ethics approval and consent to participate

The access, storage and usage of the human genetic data (POPRES and CONVERGE) were approved by the School of Biology Ethics Committee, University of St Andrews.

References

1. Barbujani G, Excoffier LGL. The history and geography of human genetic diversity. In: Stearns, Stephen C. (Ed.). *Evolution in health and disease*. Oxford: Oxford University Press, 1999. https://archive-ouverte.unige.ch/unige:93149.
2. Manica A, Prugnolle F, Balloux F. Geography is a better determinant of human genetic differentiation than ethnicity. *Hum Genet* 2005;**118**:366–71.
3. Labonte R, Polanyi M, Muhajarine N, *et al.* Beyond the divides: towards critical population health research. *Crit Public Health* 2005;**15**:5–17.
4. Parsons T. *Societies: Evolutionary and Comparative Perspectives*. Englewood Cliffs, NJ: Prentice-Hall, 1966.
5. Root M. How we divide the world. *Philos Sci* 2000;**67**:S628–39.
6. Serre D, Pääbo S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res* 2004;**14**:1679–85.
7. Rosenberg NA, Mahajan S, Ramachandran S, *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* 2005;**1**:e70.
8. Frantz A, Cellina S, Krier A, *et al.* Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *J Appl Ecol* 2009;**46**:493–505.
9. Perez MF, Franco FF, Bombonato JR, *et al.* Assessing population structure in the face of isolation by distance: are we neglecting the problem? *Divers Distrib* 2018;**24**:1883–9.
10. Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. *Curr Biol* 2005;**15**:R159–60.
11. Novembre J, Johnson T, Bryc K, *et al.* Genes mirror geography within Europe. *Nature* 2008;**456**:98–101.
12. Peter BM, Petkova D, Novembre J. Genetic landscapes reveal how human genetic diversity aligns with geography. *Mol Biol Evol* 2020;**37**:943–51.
13. Chiang CW, Mangul S, Robles C, *et al.* A comprehensive map of genetic variation in the world's largest ethnic group—Han Chinese. *Mol Biol Evol* 2018;**35**:2736–50.
14. Yang W-Y, Novembre J, Eskin E, *et al.* A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* 2012;**44**:725.
15. Yang W-Y, Platt A, Chiang CW-K, *et al.* Spatial localization of recent ancestors for admixed individuals. *G3* 2014;**4**:2505–18.
16. Coop G, Pickrell JK, Novembre J, *et al.* The role of geography in human adaptation. *PLoS Genet* 2009;**5**:e1000500.
17. Sloan CD, Duell EJ, Shi X, *et al.* Ecogeographic genetic epidemiology. *Genet Epidemiol* 2009;**33**(4):281–9.
18. Locke AE, Steinberg KM, Chiang CW, *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* 2019;**572**:323–8.
19. Galinsky KJ, Loh P-R, Mallick S, *et al.* Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. *Am J Hum Genet* 2016;**99**:1130–9.
20. McVean G. A genealogical interpretation of principal components analysis. *PLoS Genet* 2009;**5**:e1000686.
21. Cavalli-Sforza LL, Cavalli-Sforza L, Menozzi P, *et al. The History and Geography of Human Genes*. Princeton, NJ, USA: Princeton University Press, 1994.
22. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;**2**:e190.
23. Wang C-C, Yeh H-Y, Popov AN, *et al.* Genomic insights into the formation of human populations in East Asia. *Nature* 2021;591(7850):413–419.
24. Yang MA, Fan X, Sun B, *et al.* Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* 2020;**369**:282–8.
25. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet* 2019;**15**:e1008432.
26. Alanis-Lobato G, Cannistraci CV, Eriksson A, *et al.* Highlighting nonlinear patterns in population genetics datasets. *Sci Rep* 2015;**5**:8140.
27. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 2008;**40**:646–9.
28. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* 2010;**11**:94.
29. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen* 1936;**7**:179–88.
30. Deperi SI, Tagliotti ME, Bedogni MC, *et al.* Discriminant analysis of principal components and pedigree assessment of genetic diversity and population structure in a tetraploid potato panel using SNPs. *PloS One* 2018;**13**:e0194398.
31. Morrison DG. On the interpretation of discriminant analysis. *J Market Res* 1969;**6**:156–63.
32. Sugiyama M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J Mach Learn Res* 2007;**8**:1027–61.
33. Sugiyama M. Local fisher discriminant analysis for supervised dimensionality reduction. In: *Proceedings of the 23rd International Conference on Machine Learning*. Association for Computing Machinery, New York, NY, United States, 2006, p. 905–12.
34. Luo D, Liu A. Kernel Fisher discriminant analysis based on a regularized method for multiclassification and application in lithological identification. *Math Probl Eng* 2015;**2015**:1–8.
35. Weston J, Schölkopf B, Eskin E, *et al.* Dealing with large diagonals in kernel matrices. *Ann Inst Statist Math* 2003;**55**:391–408.
36. Vapnik V. The support vector method of function estimation. In: Suykens, J.A.K., Vandewalle, J. (eds) *Nonlinear Model*. Springer, Boston, MA. 1998. https://doi.org/10.1007/978-1-4615-5703-6_3.
37. Babaud J, Witkin AP, Baudin M, *et al.* Uniqueness of the Gaussian kernel for scale-space filtering. *IEEE Trans Pattern Anal Mach Intell* 1986;26–33.
38. Zelnik-Manor L, Perona P. Self-tuning spectral clustering. *Adv Neural Inf Process Syst* 2004;**17**:1601–8.
39. Attali J-G, Pagés G. Approximations of functions by a multilayer perceptron: a new approach. *Neural Netw* 1997;**10**:1069–81.
40. Baker MR, Patil RB. Universal approximation theorem for interval neural networks. *Reliab Comput* 1998;**4**:235–9.

41. Garson DG. Interpreting neural network connection weights. *Artif Intell Expert* 1991;**6**:46–51.
42. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw* 1989;**2**:359–66.
43. Miikkulainen R, Liang J, Meyerson E, *et al.* Evolving deep neural networks. In: *Artificial intelligence in the age of neural networks and brain computing* Academic Press. Cambridge, Massachusetts, 2019;293–312.
44. Nakayama K, Hirano A, Ido I. A multilayer neural network with nonlinear inputs and trainable activation functions: structure and simultaneous learning algorithm. *In IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)* IEEE. 1999;**3**:1657–61.
45. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015;**61**:85–117.
46. Excoffier L, Dupanloup I, Huerta-Sánchez E, *et al.* Robust demographic inference from genomic and SNP data. *PLoS Genet* 2013;**9**:e1003905.
47. Excoffier L, Foll M. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 2011;**27**:1332–4.
48. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. *Genetics* 2000;**156**:297–304.
49. Sanjuán R, Nebot MR, Chirico N, *et al.* Viral mutation rates. *J Virol* 2010;**84**:9733–48.
50. Nishant KT, Singh ND, Alani E. Genomic mutation rates: what high-throughput methods can tell us. *Bioessays* 2009;**31**:912–20.
51. Condit R, Levin BR. The evolution of plasmids carrying multiple resistance genes: the role of segregation, transposition, and homologous recombination. *Am Nat* 1990;**135**:573–96.
52. Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *Elife* 2021;**10**:e65366.
53. Maxwell CS, Mattox K, Turissini DA, *et al.* Gene exchange between two divergent species of the fungal human pathogen, *Coccidioides*. *Evolution* 2019;**73**:42–58.
54. Mills LS, Allendorf FW. The one-migrant-per-generation rule in conservation and management. *Conserv Biol* 1996;**10**:1509–18.
55. Ripley B, Venables B, Bates DM, *et al.* Package 'mass'. *Cran R* 2013;**538**:113–120.
56. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 2008;**24**:1403–5.
57. Nelson MR, Bryc K, King KS, *et al.* The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 2008;**83**:347–58.
58. Cai N, Bigdeli TB, Kretzschmar W, *et al.* Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* 2015;**523**:588–91.
59. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;**22**:276–82.
60. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. *Nat Genet* 2008;**40**:491–2.
61. Schwender H. Imputing missing genotypes with weighted k nearest neighbors. *J Toxicol Environ Health A* 2012;**75**:438–46.
62. Money D, Gardner K, Migicovsky Z, *et al.* LinkImpute: fast and accurate genotype imputation for nonmodel organisms. *G3* 2015;**5**:2383–90.
63. Roberts A, McMillan L, Wang W, *et al.* Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* 2007;**23**: i401–7.
64. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;**155**: 945–59.
65. Wilkinson S, Haley C, Alderson L, *et al.* An empirical assessment of individual-based population genetic statistical techniques: application to British pig breeds. *Heredity* 2011;**106**: 261–9.
66. Guinand B, Topchy A, Page K, *et al.* Comparisons of likelihood and machine learning methods of individual classification. *J Hered* 2002;**93**:260–9.
67. Cornuet J-M, Aulagnier S, Lek S, *et al.* Classifying individuals among infra-specific taxa using microsatellite data and neural networks. *Comptes rendus de l'Academie des sciences Serie III, Sciences de la vie* 1996;**319**:1167–77.
68. Battey CJ, Ralph PL, Kern AD. Predicting geographic location from genetic variation with deep neural networks. *Elife* 2020;**9**:e54507.
69. Luu K, Bazin E, Blum MG. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol Ecol Resour* 2017;**17**:67–77.
70. Frichot E, Schoville SD, Bouchard G, *et al.* Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol* 2013;**30**:1687–99.
71. Bush WS, Moore JH. Chapter 11: genome-wide association studies. *PLoS Comput Biol* 2012;**8**:e1002822.
72. Goldstein BA, Hubbard AE, Cutler A, *et al.* An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet* 2010;**11**: 1–13.
73. Sohn A, Olson RS, Moore JH. Toward the automated analysis of complex diseases in genome-wide association studies using genetic programming. In: *Proceedings of the Genetic and Evolutionary Computation Conference*. Association for Computing Machinery, New York, NY, United States, 2017, p. 489–96.
74. Qin X, Chiang CWK, Gaggiotti OE. Deciphering signatures of natural selection via deep learning. *bioRxiv* 2005;**2021**(2021):2027, 445973.
75. Taroni JN, Grayson PC, Hu Q, *et al.* MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell Systems* 2019;**8**:380, e384–94.
76. Wheeler NE, Gardner PP, Barquist L. Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLoS Genet* 2018;**14**:e1007333.
77. Mieth B, Rozier A, Rodriguez JA, *et al.* DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies. *NAR Genomics and Bioinformatics* 2021;**3**:lqab065.