# Kernel Local Fisher Discriminant Analysis of Principal Components (KLFDAPC) significantly improves the accuracy of predicting geographic origin of individuals

Xinghu Qin[1#], Charleston W. K. Chiang[2], Oscar E. Gaggiotti[1#]

[1] Centre for Biological Diversity, Sir Harold Mitchell Building, University of St Andrews, Fife, KY16 9TF, UK

[2] Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine & Department of Quantitative and Computational Biology, University of Southern California, USA

# Correspondence to: xq5@st-andrews.ac.uk and 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 geographic patterns from genetic data is principal components analysis (PCA) or a hybrid linear discriminant analysis of principal components (DAPC). However, genetic features produced from both approaches are only linear combinations of genotypes, which ineluctably miss nonlinear patterns hidden in the genetic variations and could fail to characterize the correct population structure for more complex cases. In this study, we introduce Kernel Local Fisher Discriminant Analysis of Principal Components (KLFDAPC), a nonlinear approach for inferring individual geographic genetic structure that could rectify the limitations of these linear approaches by preserving the nonlinear information and the multimodal space of samples. We tested the power of KLFDAPC to infer population structure and to predict individual geographic origin using simulations and real data sets. Simulation results showed that KLFDAPC significantly improved the population separability compared with PCA and DAPC. The application to POPRES and CONVERGE 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 (Barbujani and Excoffier 1999; Manica, et al. 2005), social stratification (Parsons 1966; Labonte, et al. 2005), as well as culture (Root 2000). For a long time, an interesting debate in population genetics is whether continuous clines or discrete clusters can better characterize human genetic variation (Serre and Pääbo 2004; Rosenberg, et al. 2005; Frantz, et al. 2009; Perez, et al. 2018). 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 fixation $F_{ST}$ increases with increasing geographic distance (Prugnolle, et al. 2005). 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 components analysis (PCA) (Novembre, et al. 2008; Chiang, et al. 2018; Peter, et al. 2020) and model-based analyses (Yang, et al. 2012; Yang, et al. 2014). Some studies have reported that the alleles showing geographic patterns are responsible for local adaptation (Hoffmann and Sgro 2011). On the other hand, alleles underlying human complex diseases such as cancer, schizophrenia, and heart disease also exhibit geographic patterns (Sloan, et al. 2009; Locke, et al. 2019). 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 individual-based tools have been developed to infer the individual geographic origin from genomic data. PCA is one of the most widely used approaches for identifying population structure and detecting loci under selection (Galinsky, et al. 2016). The link with population structure was demonstrated by (McVean 2009), who showed that the proportion of the variance explained by the first principal component computed from the genome-wide SNP genotype matrix is equal to $F_{ST}$ . Furthermore, the linear superposition of principal component (PC) maps has been used to infer the human geographic origin for various present-day and ancient individuals (Novembre, et al. 2008; Elhaik, et al. 2014; Locke, et al. 2019; Yang, et al. 2020; Wang, et al. 2021). 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 nonlinear relationship (Alanis-Lobato, et al. 2015; Diaz-Papkovich, et al. 2019). It

has also been suggested that it might be more robust to use nonlinear functions of the first two principal components, rather than higher PCs, to capture nonlinear spatial trends (Novembre and Stephens 2008).

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 (Jombart, et al. 2010). In contrast, Fisher Discriminant Analysis (LDA) (Fisher 1936) can maximize between-group variance while simultaneously minimizing within-group variance. DAPC (Discriminant Analysis of Principal Components) is a hybrid statistical technique for dimensionality reduction that combines LDA and PCA but outperforms each method in many aspects (Jombart, et al. 2010). DAPC is statistically validated for linear inference and has been successfully applied to study population structure (Jombart, et al. 2010; Deperi, et al. 2018). Nevertheless, due to the limitation of linearity, it may fail to represent data having multimodal distributions and nonlinear relations (Sugiyama 2007).

**New Approaches**

To overcome this weakness, we propose to use the kernelized and localized versions of LDA, Kernel Local Fisher Discriminant Analysis (KLFDA) (Sugiyama 2007; Luo and Liu 2015), to complement the linear discriminant analysis in population genetic studies. Kernel Local Fisher Discriminant Analysis of Principal Components (KLFDAPC) takes advantage of the nonlinear feature extraction of KLFDA and dimension reduction of PCA. KLFDAPC not only keeps the global structure of individuals, but also preserves the local multimodality within groups, and considers the nonlinear association between individuals (Sugiyama 2006, 2007). Thus, it can potentially rectify the limitations of PCA and DAPC.

In this study, we investigated the performance of KLFDAPC in population differentiation and individual geography prediction by applying the method to both simulated and empirical datasets. We simulated spatially-structured demes to test the discriminatory power of PCA, DAPC, and KLFDAPC in population structure identification. We then assessed the performance of these three approaches for predicting geographic origin of individuals using POPRES (Nelson, et al. 2008) and CONVERGE (Cai, et al. 2015) datasets. The implementation of our method is freely available in the R package *KLFDAPC* at https://xinghuq.github.io/KLFDAPC/.

## Results

### Analyses of simulated spatial scenarios

KLFDAPC is a nonlinear approach that can preserve the within-population multimodal structure properly. In general, PCA tries to capture the global variance of the data, KLFDA tries to capture both the nonlinear local variance and the between-group variance. KLFDAPC makes use of the nonlinear feature extraction of KLFDA and dimension reduction of PCA, explicitly overcoming the inefficiency of LDA and kernel transformation in processing multidimensional data (see *Methods* section below). KLFDAPC is implemented via four steps, including (1) extracting the overall variance of the genotypes using principal component analysis; (2) computing the kernel matrix **M** and affinity matrix **A**; (3) defining the local Fisher transformation matrix $\mathbf{T}_{LFDA}$ in terms of within- and between-population scatter matrices, $\bar{\mathbf{S}}^{(w)}$ and $\bar{\mathbf{S}}^{(b)}$; (4) computing an analytical form of the transformation matrix by solving a generalised eigenvalue problem. The details about these four steps are described in *Methods*.

We assessed the discriminatory power of KLFDAPC for population delineation using simulated data of spatially-structured scenarios. We simulated genome-wide SNP data under four spatial scenarios (*Methods* and Fig. S1), including two non-hierarchical spatial models (the classic island model Fig. S1A, and stepping-stone model Fig. S1C) and two hierarchical spatial models (the hierarchical island model Fig. S1B, and hierarchical stepping-stone model Fig. S1D). In each spatial model, we sampled 16 populations with a total of 3200 individuals (200 individuals per population). The hierarchical island model contains 4 regions with each region having 4 local populations. There are two regions in the hierarchical stepping model with each region composed of 8 local populations (*Methods*). We first carried out PCA on the genotype matrix of 3200 sampled individual under all simulated scenarios. We then retained the first 20 PCs to conduct DAPC and KLFDAPC using the population labels as the dependent variable. The accumulated proportion of variance for the first 20 PCs under each scenario is presented in Table S1. We projected the reduced genetic features of these four spatial scenarios onto 2D plots and 3D plots as the representation of population genetic structure (Fig. 1 & Fig. S2). We also tested the predictive power of the three approaches to assign the individuals to the sampling populations. The performance of KLFDAPC was evaluated using PCA and DAPC as the benchmarks.

We can easily identify the four spatial scenarios (discriminate between scenarios) from their genetic structures projected by PCA (Fig. 1A-D), DAPC (Fig. 1E-H) and KLFDAPC (Fig. 1I-L), in which genetically closer populations tended to aggregate together. All methods successfully discriminated the regions under the hierarchical spatial scenarios (hierarchical island model, Fig. 1B, F, J, and hierarchical stepping-stone model, Fig. 1D, H, L). However, PCA and DAPC both failed to clearly delineate local populations under all four scenarios (Fig. 1A-H). In contrast, KLFDAPC features clearly distinguished the stratification among local populations, especially for the stepping-stone and hierarchical stepping-stone models (Fig. 1K, L & Fig. S2). 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 (classical island model and hierarchical island model), where populations within regions tend to overlap (Fig. 1). Note that including the third reduced feature provided more detailed structure representation and allowed populations within region 2 under the hierarchical stepping-stone model to be clearly identified (Fig. S2). More precisely, the first reduced feature well discriminated between the two higher level regions in the hierarchy while the second and third reduced feature together discriminated among populations within groups.

Consistent with the graphical representation of the spatial structures, the discriminatory accuracy and Cohen's Kappa coefficient ($\kappa$) of KLFDAPC were much higher than those achieved by PCA and DAPC under all spatially-structured scenarios except the hierarchical stepping-stone model (Fig. S3). Nevertheless, we observed that the performance of KLFDAPC was only slightly better than that of DAPC under the hierarchical stepping-stone model if only the first two reduced features were used as the predictor variables (Fig. S3). We therefore recommend considering more features (i.e., the first three features in Fig. S2) to fully characterise population structure under complex scenarios.

As Fig. S4 illustrates, the pattern of local genetic aggregation is potentially sensitive to the parameter $\sigma$ in KLFDAPC. $\sigma$ adjusts the dispersal/aggregation of the local structure, thus changing the geometry (shape) of the stratification constituted by the first two reduced features. Lowering the $\sigma$ values of KLFDAPC could increase the discrimination of discrete clusters, thus increasing the ability of KLFDAPC to delineate distinct aggregates (Fig. S4).

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., $\sigma$ in Gaussian kernel). Therefore, we recommend users to vary kernel parameter values to explore how it influences the results.

**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 principal components 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 (Figs. 2A, D; Table 1, PC1 vs. longitude, R= 0.872, PC2 vs. latitude, R=0.873), as previously reported (Novembre, et al. 2008). Applying DAPC and KLFDAPC to the top 20 PCs also clearly captured a similar correlation with the geographic map (Fig. 2B, C, E, F). Notably, DAPC and KLFDAPC rectified the genetic distance between populations from geographically close/distant countries compared with PCA. 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 keeping PL (Poland) and RU (Russia) distant from each other, thus significantly improving the alignment between the genetic distance and the geographic distance (Fig. 2; Table 3, PC1 vs. longitude, R= 0.872; KLFDAPC1 vs. longitude, R= 0.886; PC2 vs. latitude, R= 0.873; KLFDAPC2 vs. latitude, R= 0.934, the difference in $R^2$ between PCA and KLFDAPC was significant as indicated in Table 3). One advantage of KLFDAPC is that $\sigma$ can be varied to adjust the global/local scaling of the genetic structure. Increasing $\sigma$ 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 (Fig. S5). KLFDAPC with $\sigma$ =5 achieved the best in predicting the geographic locations of POPRES individuals compared to PCA and DAPC (Table 2-3). Overall, lower $\sigma$ values tend to aggregate individuals into compact clusters, while high $\sigma$ values make the individuals more scattered.

Summary statistics measuring the performance of the three approaches in inferring individual geographic locations are presented in Tables 1-3. Of note, by adjusting the value of $\sigma$, KLFDAPC ($\sigma$ = 2.5) could outperform PCA and DAPC in geometric (shape) similarity tested by the Procrustes analysis (Table 1). According to the Procrustes statistics, KLFDAPC with $\sigma$ = 1 outperformed both PCA and DAPC based on the association between the first two reduced features and geographic locations after Procrustes transformation (Table 1). When

predicting the individual origin via a MLP model using the first two reduced features as the predictor variables, KLFDAPC with $\sigma$ =5 performed the best among the three methods (Table 2). DAPC and PCA showed similar power in predicting the individual geographic locations (Table 3). In contrast, KLFDAPC ($\sigma$ =5) did significantly better than PCA and DAPC (Table 3, Fig. S5-S6).

**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 CONVERGE data consists 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 (Chiang, et al. 2018).

As the benchmark, PCA performed poorly in recapitulating the geography of individuals, as PC1 corresponded poorly to both latitude and longitude (Table 4, Fig. 3A, D; PC1 vs. longitude R= 0.0508, PC1 vs. latitude R= -0.351). However, we can still observe a significant North-South gradient (Table 4, Fig. 3A & S; PC2 vs. latitude R=0.6404). Compared to PCA, the reduced features obtained from DAPC better represented the genetic gradients along longitude and latitude, and significantly aligned with East China on the map, where most of the participants were born (Fig. 3B, E; Table 4). 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 longitude (East-West axis) and latitude (South-North axis), and aligned significantly better with geography (KLFDAPC2 vs. latitude R=0.7357, Table 4, Fig. 3 & Fig. S7-S8). Likewise, increasing $\sigma$ values increased the spatial genetic clines along longitude and latitude (Fig. S7). However, increasing $\sigma$ values may make the populations indiscernible, especially for populations that are both genetically and geographically related, such as Shanghai, Jiangsu and Zhejiang (Fig. S7).

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. All methods failed to predict individual longitude using neural networks (Table 5). However, they still performed well in predicting the individual latitude (Tables 4-6). PCA did worst in predicting the individual latitude among these three approaches ($R^2$=0.727). 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 4-6).

Beside the correlation between the first reduced genetic feature (RD1) and the longitude as reported in tables 4-5, we also found that the first two genetic features of all three approaches were significantly geographically clustered as indicated by Moran's $I$ (Fig. S9). Notably, genetic features obtained from KLFDAPC ($\sigma$=5, 1, 2.5) had higher Moran's $I$ than those obtained from DAPC and PCA (Fig. S9). In addition, various per capita economic metrics of the provinces also displayed significant spatial pattern by Moran's $I$ (Fig. S9, $p$ value=0.0013, 9.81E-05, 8.18E-06 respectively for GDP increase, average annual GDP, and median annual GDP). Compared to previously reported GDP map of China (He, et al. 2014; Xiang, et al. 2020), there was an apparent association between the first reduced genetic feature of KLFDAPC and the provincial economic development metrics that would not be observed if we used PCs, although the association is not significant after correcting for multiple testing (Fig. S10).

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 populations using CONVERGE dataset (Tables 4-6, Procrustes correlations, linear correlations with latitude and longitude, predicting the individual locations), suggesting that genetic features produced by KLFDAPC may be a better surrogate of geographic coordinates than PCs.

## Discussion

In this study, we introduced KLFDAPC, a nonlinear approach for inferring population genetic structure and individual geographic origin. We showed that KLFDAPC outperforms both PCA and DAPC in population structure discrimination and in predicting individual geographic origin using simulated spatially-structured scenarios and empirical population datasets. In addition, KLFDAPC also explores population structure at different spatial scales by varying a kernel parameter $\sigma$. Larger $\sigma$ values provide a more continuous representation of genetic variation while small values lead to more discrete representations. Analytical results of POPRES data revealed that populations within the European continent presented significant genetic clines corresponding to geography. All three methods retrieved a strong correspondence between genetic structure and geography. Despite the high correlation

between reduced genetic features and geography in PCA and DAPC, KLFDAPC still out-performed both of them in inferring the individual geographic locations based on four different metrics (genotype map projection, linear correlation, Procrustes analysis, feature predictive performance; Table 1-3, Fig. 2). Moreover, KLFDAPC could capture several discrete genetic clusters with a low $\sigma$ (Figs. S5& S7). 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 (Figs. 3 & S8). Again, KLFADPC outperformed both PCA and DAPC in predicting the individual geography (of Chinese populations) according to the four statistical metrics (Table 4-6, Fig. 3). 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 nonlinear genetic features but also preserves the multimodal structure within populations compared to PCA and DAPC, could be helpful for many population genetic studies.

Model-based approaches generally require users to define a nonlinear function modelling the slope of allele frequency in geographic locations (Yang, et al. 2012; Yang, et al. 2014). Any pre-specified parametric function is unlikely to sufficiently capture complicated relations between SNPs and geographic space, such as multiple modes or peaks in the allele frequency surface as reported by (Yang, et al. 2012). In contrast, KLFDAPC is a model-free method that captures both nonlinear genetic information between individuals and multimodal structure within populations. It does not require a nonlinear function being defined, but only needs to choose an appropriate kernel. KLFDAPC users can adjust the scatter or aggregation (continuous or discrete) of the population structure by changing the kernel transformation parameter, for example, $\sigma$ in the Gaussian kernel used in our study. Our simulation study and real data analyses showed that larger $\sigma$ tends to produce more scattered or continuous clines, and vice versa (Fig. S4). In this way, users could explore the desired spatial genetic patterns, either continuous or discrete, depending on research objectives. By tuning $\sigma$, it is possible to maximise 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 CONVERGE data, PCA tended to provide a poor indication of the correct geographic origin of individuals. A previous study of the population structure in

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 (Chiang, et al. 2018). In our present analyses, 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 (Roberts, et al. 2007; Schwender 2012; Money, et al. 2015). Note that KLFDAPC uses the *k*-nearest neighbour algorithm 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 nonlinear 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 (Pritchard, et al. 2000) and could mask important evolutionary processes such as migration and cross-breeding (Wilkinson, et al. 2011). LDA suffers from this problem as it minimizes the within-group variations based on group means, and can thus miss admixed or hybrid individuals as well as important evolutionary processes, leading to inaccurate representations of genetic structure. Unlike LDA and other within-group minimizing/averaging approaches, KLFDAPC is less affected by pre-grouping. The main reason is that KLFDAPC computes the within- and between- population affinity based on the *k*-nearest neighbour algorithm, minimising the influence of 'local outliers' (i.e., migrants) on the inference of population structure. Meanwhile, outlier genotypes are down weighted by local affinity instead of the group mean used by LDA. Therefore, KLFDAPC can well preserve the multimodal genetic structure within populations and potentially mitigates the problems and concerns caused by group-mean (or group-centroid) based method.

Besides predicting the individual geographic origin, we also provided an implementation of genome scan to identify genomic regions involved in local adaptation based on KLFDAPC (https://xinghuq.github.io/KLFDAPC/articles/Genome_scan_KLFDAPC.html). Many geographic or environmental selection gradients are non-linear, and many complex

phenotypes are not linearly correlated to genotypes too. We expect that KLFDAPC would be informative when applied to certain designs of population genetic studies. First, population structure integrating nonlinear information between individuals could correct spurious results in inferring individual geographic origin (i.e., Han Chinese example in our study) and rectify genetic signatures related to nonlinear spatial or environmental clines. The nonlinear genetic structure with a nonlinear outlier identification model (as opposed to linear regression models in *pcadapt* (Luu, et al. 2017) and LFMM (Frichot, et al. 2013)) 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 adaptation in the same way as *pcadapt* (Luu, et al. 2017) does. Second, nonlinear features of KLFDAPC could correct for stratification in genome-wide association studies, in combination with GWAS, potentially minimizing spurious associations while also maximizing detection power.

Machine learning algorithms have enabled us to capture the systematic signatures of biological or genetic patterns of samples, allowing for the association of genes to phenotypes/diseases and facilitating molecular-based medical applications. KLFDAPC will be a necessary tool that recapitulates genetic structure and provides new complementary insights into signatures of evolutionary events. This approach is potentially also applicable to other Omics data throughout the biological sciences, including applications in medicine and agriculture.

## Materials and Methods

### Kernel Local Fisher Discriminant Analysis of Principal Components (KLFDAPC)

Kernel Local Fisher discriminant analysis (KLFDA) is a nonlinear local linear discriminant analysis implemented with the kernel transformation technique (Sugiyama 2007; Luo and Liu 2015). KLFDA integrates locality-preserving projection (LPP) and Fisher discriminant analysis (FDA). It not only maximizes between-group separability but also preserves within-group local structure. Thus, it is good at dealing with multimodal data, such as samples from multiple sources (populations).

KLFDA works well for low dimensional data. However, the kernel transformation faces two key problems, the heavy computation cost and large diagonals in the kernel matrix if the

number of variables is much larger than the number of samples ($p >> n$)(Weston, et al. 2003), which is typically the case for genetic data with millions of SNP genotypes. One solution is to use low dimensional features but also capture the largest variance of the original data. Kernel Local Fisher Discriminant Analysis of Principal Components (KLFDAPC) integrates dimension reduction of PCA and the nonlinear feature extraction of KLFDA, making the KLFDAPC more 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, species counts. In what follows, we describe the rationale for this approach and the steps required to implement it in general.

**Principal component analysis of genomic data**

The first step of implementing KLFDAPC is to extract the overall variance of the genotypes using principal component analysis. Let $G$ denote the $n \times L$ genotype matrix with $n$ individuals in rows and $L$ loci/markers in columns coded by the number of derived alleles $0, 1, 2$. The individuals are labelled by $i = (1, \ldots, n)$ with collected population labels $y_i = (1, \ldots, c)$. In the principal component analysis, we ignore the population labels.

Here we used the PCA implementation according to (Patterson, et al. 2006). Let $g(i,l)$ be the genotype (number of reference alleles) of individual $i$ at locus $l$. The mean number of reference alleles per individual for locus $l$ is $\mathrm{u(l)} = \frac{\sum_{i=1}^{n} g(i,l)}{n}$. Thus, the scaled copy number of the reference allele is $g(i,l) - u(l)$ . If we note the frequency of the reference allele by $p_l = \mathrm{u(l)}/2$, the $(i, l)$-th entry of the normalized matrix $\mathbf{G}$ is,

$$G_{i,l} = \frac{g_{i,l} - u(l)}{\sqrt{p_l(1-p_l)}}. \qquad (1)$$

The covariance of the genotype matrix is defined by,

$$\mathbf{Cov} = \frac{\mathbf{G}\mathbf{G}'}{L}. \qquad (2)$$

**Cov** is a $n$ ⊠ $n$ matrix with individuals $i$ and $j$ as the entry. Based on the above definition, the $(i, j)$-th entry of the genotype covariance matrix is,

$$Cov_{i,j} = \frac{1}{L}\sum_{l=1}^{L} \frac{(g_{i,l} - u(l))(g_{j,l} - u(l))}{p_l(1-p_l)}, \quad (3)$$

The principal components that best explain the genotypic variability amongst the *n* individuals can then be obtained through singular value decomposition (SVD) of the covariance matrix **Cov**, as explained in Patterson et al (2006).

**Kernel Local Fisher Discriminant Analysis of Principal components**

In the second step, we use the top *P* principal components to conduct the KLFDA. *P* can be determined in principal component analysis by looking at the cumulative variance of the principal components. 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 among population components. Thus, each individual has a population label $y_i \in (1, 2, \ldots, c)$. Although an individual could be a recent migrant and, therefore, the population labels might not represent its 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 $\bar{\mathbf{S}}^{(w)} \in R^n$ and the between-population scatter matrix $\bar{\mathbf{S}}^{(b)} \in R^n$, and then carry out a generalised eigenvalue decomposition. More details about the kernel local Fisher discriminant analysis formulation can be found in (Sugiyama 2006, 2007). 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 $\mathbf{T}_{LFDA}$ in terms of within- and between-population scatter matrices $\bar{\mathbf{S}}^{(w)}$ and $\bar{\mathbf{S}}^{(b)}$; and (iii) computing an analytical form of the transformation matrix by solving a generalised eigenvalue problem.

**Computing the kernel matrix**

Once the data have been compressed to *P* principal component, KLFDA first transforms the principal component scores into a kernel matrix **M** via nonlinear mapping using a kernel function (Vapnik 1998). In this study, we use a Gaussian kernel. Let $\mathbf{x}_i = (x_{ip})$ and $\mathbf{x}_j = (x_{jp})$ be vectors containing the top *P* principal components for individuals *i* and *j*. The elements of the Gaussian kernel matrix **M** can be defined as,

$$M_{i,j} = \frac{1}{\sqrt{2\pi}\,\sigma} exp\ \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \qquad (4)$$

with $\sigma$ determining the width of the Gaussian kernel (Babaud, et al. 1986). The kernel matrix can be viewed as a *n*-dimensional genetic distance matrix between pairs of individuals where each individual has a population label $y_i$ □ (*1*, *2*, …, *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 (Zelnik-Manor and Perona 2004). 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 estimated from Eq. 4. 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 (Zelnik-Manor and Perona 2004),

$$A_{i,j} = \exp\ \left(-\frac{\mathbf{m}_i - \mathbf{m}_j{}^2}{\sigma_i \sigma_j}\right), \qquad (5)$$

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

$$\sigma_i = \mathbf{m}_i - \mathbf{m}_i^K, \qquad (6)$$

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 $\boldsymbol{T_{LFDA}}$**

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 $\bar{W}_{i,j}^{(w)}$ represent the within- population affinity and $\bar{W}_{i,j}^{(b)}$ represent the between-population affinity. Let $\bar{\mathbf{S}}^m$ be the local mixture scatter matrix defined by $\bar{\mathbf{S}}^m = \bar{\mathbf{S}}^{(w)} + \bar{\mathbf{S}}^{(b)}$. The local within-population scatter matrix $\bar{\mathbf{S}}^{(w)}$ and the local between-population scatter matrix $\bar{\mathbf{S}}^{(\mathrm{b})}$ can be obtained as follows.

$$\bar{\mathbf{S}}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} \bar{W}_{i,j}^{(w)} \left(\mathbf{m}_i - \mathbf{m}_j\right)\left(\mathbf{m}_i - \mathbf{m}_j\right)^T, \qquad (7)$$

$$\bar{\mathbf{S}}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} \bar{W}_{i,j}^{(b)} \left(\mathbf{m}_i - \mathbf{m}_j\right)\left(\mathbf{m}_i - \mathbf{m}_j\right)^T, \qquad (8)$$

$$\bar{W}_{i,j}^{(w)} = \left\{ \begin{array}{ll} \frac{A_{i,j}}{n_y}, & \text{if } y_i = y_j = y \\ 0, & \text{if } y_i \neq y_j . \end{array} \right\}, \qquad (9)$$

$$\bar{W}_{i,j}^{(b)} = \left\{ \begin{array}{ll} A_{i,j}\left(\frac{1}{n} - \frac{1}{n_y}\right), & \text{if } y_i = y_j = y \\ \frac{1}{n}, & \text{if } y_i \neq y_j . \end{array} \right\}, \qquad (10)$$

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

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

$$\bar{\mathbf{S}}^{(m)} = \frac{1}{2}\sum_{i,j=1}^{n} \bar{W}_{i,j}^{(m)} \left(\mathbf{m}_i - \mathbf{m}_j\right)\left(\mathbf{m}_i - \mathbf{m}_j\right)^T, \qquad (11)$$

$$\bar{W}_{i,j}^{(m)} \equiv \left\{ \begin{array}{ll} \frac{A_{i,j}}{n}, & \text{if } y_i = y_j \\ \frac{1}{n}, & \text{if } y_i \neq y_j . \end{array} \right\}, \qquad (12)$$

Therefore,

$$\bar{\mathbf{S}}^{(m)} = \frac{1}{2}\sum_{i,j=1}^{n} \bar{W}_{i,j}^{(m)}\left(\mathbf{m}_i\mathbf{m}_i^T + \mathbf{m}_j\mathbf{m}_j^T - \mathbf{m}_i\mathbf{m}_j^T - \mathbf{m}_j\mathbf{m}_i^T\right)$$

$$= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} \bar{W}_{i,j}^{(m)}\right)\mathbf{m}_i\mathbf{m}_i^T - \sum_{i,j=1}^{n} \bar{W}_{i,j}^{(m)}\,\mathbf{m}_i\mathbf{m}_j^T. \qquad (13)$$

Eq. (13) can be expressed in matrix form as

$$\bar{\mathbf{S}}^{m} = \mathbf{M}\bar{\mathbf{L}}^{(m)}\mathbf{M}^T, \qquad (14)$$

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

Using $\bar{\mathbf{S}}^{(w)}$ and $\bar{\mathbf{S}}^{(b)}$ , the local Fisher transformation matrix $\mathbf{T}_{LFDA}$ can be defined as (Sugiyama 2006),

$$\mathbf{T}_{LFDA} = \underset{T \in R^{d\times m}}{argmax} \left[ tr\left( \left( \mathbf{T}^T \bar{\mathbf{S}}^{(w)} \mathbf{T} \right)^{-1} \mathbf{T}^T \bar{\mathbf{S}}^{(b)} \mathbf{T} \right) \right], \qquad (15)$$

$T_{LFDA}$ is the ratio of between-population ($\bar{S}^{(b)}$ ) and within-population ($\bar{S}^{(w)}$) variances, also known as the *F*- statistic in LDA, which is used to find the best transformation matrix to maximize Fisher's criterion (Fisher 1936).

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

Noting that $\mathbf{M}$ is a symmetric matrix, $\mathbf{T}_{LFDA}$ is obtained by solving (Sugiyama 2006, 2007),

$$\mathbf{M}\bar{\mathbf{L}}^{(b)}\mathbf{M}\varphi = \lambda \mathbf{M}\bar{\mathbf{L}}^{(w)}\mathbf{M}\varphi, \qquad (16)$$

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

In practice, Eq. (16) cannot be solved because $\bar{\mathbf{L}}^{(w)}$ is always singular. Therefore, Sugiyama (2007) proposes regularising $\mathbf{M}\bar{\mathbf{L}}^{(w)}\mathbf{M}$ and solving instead.

$$\mathbf{M}\bar{\mathbf{L}}^{(b)}\mathbf{M}\varphi = \lambda\left(\mathbf{M}\bar{\mathbf{L}}^{(w)}\mathbf{M} + \varepsilon \mathbf{I}_n\right)\varphi, \qquad (17)$$

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

**Simulations**

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 (Excoffier and Foll 2011; Excoffier, et al. 2013).

For each model, we simulated 16 populations comprising of 2000 haploid individuals (equivalent to 1000 diploid individuals). The island model (Fig. S1A) (including hierarchical island model, Fig. S1B) and stepping stone model (Fig. S1C) (including hierarchical stepping stone model, Fig. S1D) differ in the composition of aggregates (regions) and migration

pattern. The hierarchical island model consists of four regions with each region comprising 4 populations. The hierarchical stepping-stone model consists of two regions with each region comprising 8 populations. We simulated 44 independent chromosomes with 100 Kb DNA sequences per chromosome assuming a finite mutation model with a constant mutation rate of $u=1\times10^{-8}$ per bp per generation and a recombination rate of $r=1\times10^{-8}$ per bp per generation for all scenarios. In the case of the non-hierarchical scenarios (Island model and stepping-stone model), we assumed the migration rates between populations to be 0.001. In the case of the hierarchical models (hierarchical island model and hierarchical stepping-stone model), migration rates between pairs of populations within regions were 0.001 and migration between populations from different regions were 0.0001. The four scenarios and the respective parameter values used in the simulations are presented in Table S1.

We carried out 10 independent simulations for each scenario and sampled 200 individuals from each population. In total, we obtained 3,200 individuals from 16 populations under each spatial scenario. Each scenario generated more than 27,000 polymorphic sites. We randomly selected 10,000 sites (biallelic) per scenario for downstream analysis.

We removed monomorphic SNPs and filtered the SNPs with a MAF > 0.05 for the simulated data. For each scenario, we first carried out a principal component analysis on the genotype matrix. We then used the first 20 PCs as the inputs to conduct DAPC and KLFDAPC. PCA was implemented in R using the *prcomp* function from *stats* package with the genotype matrix being normalized before analysis (R Core Team 2013). DAPC was implemented using the first 20 PCs via *lda* function from *MASS* package (Ripley, et al. 2013), which is initially employed by *dapc* function in *adegenet* package (Jombart 2008). We used the radial basis function kernel (RBF kernel), also known as Gaussian kernel to carry out KLFDAPC using the same 20 PCs. DAPC and KLFDAPC were conducted using the source population names as the group labels. In terms of KLFDAPC, parameter $\sigma$ determines the width of the Gaussian kernel and therefore, determine the delineation of populations. To investigate the effect of $\sigma$ on population discrimination, we varied the $\sigma$ values from 0.5- 5 to find the best feature representation.

**Testing the performance of different approaches in identifying population structure using simulated SNP data**

We tested the performance of our method to discriminate populations under four spatial scenarios using the simulated SNP datasets. We presented the population structure (reduced genetic features) produced from three approaches and estimated the accuracy and Cohen's Kappa coefficient ($\kappa$) of these genetic features to predict the individuals to the labelled populations (Supplementary Methods, Supplementary Materials). The accuracy and Cohen's Kappa coefficient ($\kappa$) were estimated via a MLP model (Supplementary Methods, Supplementary Materials). Using PCA and DAPC as the benchmarks, we compared the differences of the accuracy and Cohen's Kappa coefficient ($\kappa$) between three approaches (Supplementary Methods, Supplementary Materials).

**Testing the performance of different approaches in predicting the geographic locations of individuals using POPRES and CONVERGE data**

We tested the performance of our approach to predict the geographic locations of individuals using two datasets, the European populations from POPRES datasets (Nelson, et al. 2008) (dbGaP accession number phs000145.v4.p2), and Han Chinese populations from CONVERGE data (Cai, et al. 2015) (http://www.ebi.ac.uk/ena/data/view/PRJNA289433). The details on data quality control can be found in Supplementary Methods (Supplementary Material). To assess the performance of three approaches (PCA, DAPC, and KLFDAPC) in inferring the geographic origin, four metrics were used in our study: (1) individual geographic location projection, (2) correlation analysis, (3) Procrustes analysis, (4) feature predictive performance comparison using neural network model resampling. Details of testing on each metric can be found in Supplementary Methods (Supplementary Material).

**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 from dbGaP database with accession number phs000145.v4.p2. The CONVERGE data is available at the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/data/view/PRJNA289433). The package KLFDAPC used for analysis is available at https://xinghuq.github.io/KLFDAPC/.

**Ethical Statement**

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.

**Acknowledgments**

XHQ was supported by CSC-University of St Andrews Joint Scholarship.

**Author contribution**

XQ and OEG designed the study. CKWC reviewed the design, acquired the datasets and provided computational resources. XQ compiled the package, carried out the analyses and interpreted results with input from CWKC and OEG. XQ wrote the manuscript with the input of CKWC and OEG. All authors contributed to editing and revisions of the manuscript.

**Conflict of interest**

The authors declare that they have no conflict of interests.

## Tables

**Table 1.** Procrustes correlations between the first two reduced features and locations for different approaches (POPRES data)

| | Procrustes Analysis | | | Correlation between the first two reduced features and locations after Procrustes transformation | | | |
|---|---|---|---|---|---|---|---|
| Methods | Procrustes Sum of Squares | Procrustes correlation | Significance | RD1 vs. Longitude | P value | RD2 vs. Latitude | P value |
| PCA | 0.297 | 0.838 | 1.00E-05 | 0.872 | 2.2E-16 | 0.873 | 2.2E-16 |
| DAPC | 0.290 | 0.843 | 1.00E-05 | 0.881 | 2.2E-16 | 0.912 | 2.2E-16 |
| KLFDAPC (σ =1) | 0.298 | 0.838 | 1.00E-05 | **0.886** | 2.2E-16 | **0.934** | 2.2E-16 |
| KLFDAPC (σ =2.5) | 0.265 | **0.857** | 1.00E-05 | 0.881 | 2.2E-16 | 0.912 | 2.2E-16 |
| KLFDAPC (σ =5) | 0.325 | 0.822 | 1.00E-05 | 0.879 | 2.2E-16 | 0.910 | 2.2E-16 |

**Table 2.** Performance of different methods to predict individual locations as measured by a MLP approach with cross-validation (POPRES data).

| | Longitude | | | Latitude | | | Correlation between observed & predicted location using the optimal MLP model | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Methods | RMSE | $R^2$ | MAE | RMSE | $R^2$ | MAE | RD1 vs. Longitude | P value | RD2 vs. 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 |

**Table 3.** The difference of $R^2$ (observed value vs. predicted value) between different models for predicting the individual longitude and latitude (POPRES data). Upper diagonal: estimates of the difference. Lower diagonal: *p*-value for H0: difference = 0.

| ***RD1 vs. longitude*** | | | | | |
|---|---|---|---|---|---|
| | PCA | DAPC | KLFDAPC_ σ _1 | KLFDAPC_ σ _2.5 | KLFDAPC_ σ _5 |
| 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 | 0.0045 | | -1.60E-02 |
| KLFDAPC_ σ _5 | **0.011** | **0.026** | 2.26E-05 | 0.084 | |
| ***RD2 vs. 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 | |

$R^2$ (true value vs. predicted value) was estimated after 5-fold cross-validations, repeated 5 times. Comparison between models was tested using a pairwise t-test between 100 resamples. *p*-value adjustment: *Bonferroni*.

**Table 4**. Procrustes correlations between the first two reduced features and individual locations (CONVERGE data)

| | Procrustes Analysis | | | Correlation between the first two PCs and locations after Procrustes transformation | | | |
|---|---|---|---|---|---|---|---|
| | Procrustes Sum of Squares | Procrustes correlations | Significance | RD1 vs. longitude | P value | RD2 vs. latitude | P value |
| PCA | 0.986 | 0.1196 | 1E-05 | 0.0508 | 2.2E-16 | 0.6404 | 2.2E-16 |
| DAPC | 0.952 | 0.2197 | 1E-05 | 0.1898 | 2.2E-16 | 0.7061 | 2.2E-16 |
| KLFDAPC_σ_0.5 | 0.946 | 0.2322 | 1E-05 | **0.2005** | 2.2E-16 | **0.7357** | 2.2E-16 |
| KLFDAPC_σ_1 | 0.953 | 0.216 | 1E-05 | 0.1893 | 2.2E-16 | 0.7242 | 2.2E-16 |
| KLFDAPC_σ_2.5 | 0.961 | 0.1978 | 1E-05 | 0.1651 | 2.2E-16 | 0.7255 | 2.2E-16 |
| KLFDAPC_σ_5 | 0.955 | 0.2134 | 1E-05 | 0.1901 | 2.2E-16 | 0.7092 | 2.2E-16 |

**Table 5**. Performance of different methods to predict individual locations as measured by a MLP approach with cross-validation (CONVERGE data)

| | Longitude | | | Latitude | | | Correlation between observed & predicted location using the optimal MLP model | | |
|---|---|---|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | MAE | RMSE | $R^2$ | MAE | RD1 vs. Longitude | RD2 vs. Latitude | P value |
| PCA | 37.637 | 0.003 | 20.310 | 5.688 | 0.530 | 4.460 | NA | 0.7274 | 2.2E-16 |
| DAPC | 35.000 | 0.038 | 16.641 | 5.682 | 0.549 | 4.460 | NA | 0.7409 | 2.2E-16 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| KLFDAPC_σ_0.5 | 36.236 | 0.018 | 18.071 | **5.615** | **0.592** | **4.398** | NA | 0.7670 | 2.2E-16 |
| KLFDAPC_σ_1 | 35.416 | 0.012 | 17.403 | 5.638 | 0.558 | 4.442 | NA | 0.7471 | 2.2E-16 |
| KLFDAPC_σ_2.5 | 34.220 | 0.006 | 16.625 | 5.650 | 0.557 | 4.457 | NA | 0.7459 | 2.2E-16 |
| KLFDAPC_σ_5 | 34.730 | 0.015 | 16.854 | 5.714 | 0.553 | 4.496 | NA | 0.7427 | 2.2E-16 |

The best statistic is marked in bold.

**Table 6**. The difference of $R^2$ (observed values vs. predicted values) between different models for predicting the individual latitude (CONVERGE Data). Upper diagonal: estimates of the difference. Lower diagonal: *p*-value for H0: difference = 0.

| | 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 | |

$R^2$ (observed values vs. predicted values) was estimated after 5-fold cross-validations, repeated 5 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.

## Figure legends

**Fig. 1**. 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 kept in DAPC and KLFDAPC analyses. The same colour in the scatter plots represents the same region. Individuals are grouped by population names.

**Fig. 2.** 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 (D-F) using three approaches. (A&D) Population geographic genetic structure (A) and individual geographic locations (D) projected by the first two PCs. (B&E) Population geographic genetic structure (B) and individual geographic locations (E) projected by the first two reduced features of DAPC. (C&F) Population geographic genetic structure (C) and individual geographic locations (F) projected by the first two reduced features of KLFDAPC, with $\sigma$ = 5. 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.

**Fig. 3.** Genetic structure of CONVERGE dataset represented by the first two reduced features from PCA (A), DAPC (B), and KLFDAPC (C) and projected individual geographic locations within China (D-F) from three approaches. A, D: Population geographic genetic structure (A) and individual geographic locations (D) projected by the first two PCs. B, E: Population geographic genetic structure (B) and individual geographic locations (E) projected by the first two reduced features of DAPC. C, F: Population geographic genetic structure (C) and individual geographic locations (F) projected by the first two reduced features of KLFDAPC with $\sigma$ = 0.5. 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; GuangxiZhuangzu, GX; Chongqing, CQ; Fujian, FJ; Gansu, GS; Jilin, JL; Anhui, AH; Hainan, HAN.

## References


Alanis-Lobato G, Cannistraci CV, Eriksson A, Manica A, Ravasi T. 2015. Highlighting nonlinear patterns in population genetics datasets. Scientific Reports 5:8140.

Babaud J, Witkin AP, Baudin M, Duda RO. 1986. Uniqueness of the Gaussian kernel for scale-space filtering. IEEE Transactions on pattern analysis and machine intelligence:26-33.

Barbujani G, Excoffier LGL. 1999. The history and geography of human genetic diversity: Oxford University Press.

Cai N, Bigdeli TB, Kretzschmar W, Li Y, Liang J, Song L, Hu J, Li Q, Jin W, Hu Z. 2015. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature 523:588-591.

Chiang CW, Mangul S, Robles C, Sankararaman S. 2018. A comprehensive map of genetic variation in the world's largest ethnic group—Han Chinese. Molecular Biology and Evolution 35:2736-2750.

Deperi SI, Tagliotti ME, Bedogni MC, Manrique-Carpintero NC, Coombs J, Zhang R, Douches D, Huarte MA. 2018. Discriminant analysis of principal components and pedigree assessment of genetic diversity and population structure in a tetraploid potato panel using SNPs. PloS one 13:e0194398.

Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. 2019. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS genetics 15:e1008432.

Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Calò CM, De Montis A, Atzori M, Marini M, Tofanelli S, Francalacci P. 2014. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nature communications 5:1-13.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet 9:e1003905.

Excoffier L, Foll M. 2011. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27:1332-1334.

Fisher RA. 1936. The use of multiple measurements in taxonomic problems. Annals of eugenics 7:179-188.

Frantz A, Cellina S, Krier A, Schley L, Burke T. 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? Journal of Applied Ecology 46:493-505.

Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. Molecular Biology and Evolution 30:1687-1699.

Galinsky KJ, Loh P-R, Mallick S, Patterson NJ, Price AL. 2016. Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. The American Journal of Human Genetics 99:1130-1139.

He Q, Bertness MD, Bruno JF, Li B, Chen G, Coverdale TC, Altieri AH, Bai J, Sun T, Pennings SC. 2014. Economic development and coastal ecosystem change in China. Scientific Reports 4:1-9.

Hoffmann AA, Sgro CM. 2011. Climate change and evolutionary adaptation. Nature 470:479-485.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11:94.

Labonte R, Polanyi M, Muhajarine N, Mcintosh T, Williams A. 2005. Beyond the divides: Towards critical population health research. Critical Public Health 15:5-17.

Locke AE, Steinberg KM, Chiang CW, Service SK, Havulinna AS, Stell L, Pirinen M, Abel HJ, Chiang CC, Fulton RS. 2019. Exome sequencing of Finnish isolates enhances rare-variant association power. Nature 572:323-328.

Luo D, Liu A. 2015. Kernel Fisher discriminant analysis based on a regularized method for multiclassification and application in lithological identification. Mathematical Problems in Engineering 2015.

Luu K, Bazin E, Blum MG. 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Molecular Ecology Resources 17:67-77.

Manica A, Prugnolle F, Balloux F. 2005. Geography is a better determinant of human genetic differentiation than ethnicity. Human Genetics 118:366-371.

McVean G. 2009. A genealogical interpretation of principal components analysis. PLoS genetics 5.

Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S. 2015. LinkImpute: fast and accurate genotype imputation for nonmodel organisms. G3: Genes, Genomes, Genetics 5:2383-2390.

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G. 2008. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. The American Journal of Human Genetics 83:347-358.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR. 2008. Genes mirror geography within Europe. Nature 456:98-101.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nature genetics 40:646-649.

Parsons T. 1966. Societies: Evolutionary and comparative perspectives: Prentice-Hall Englewood Cliffs, NJ.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet 2:e190.

Perez MF, Franco FF, Bombonato JR, Bonatelli IA, Khan G, Romeiro□Brito M, Fegies AC, Ribeiro PM, Silva GA, Moraes EM. 2018. Assessing population structure in the face of isolation by distance: Are we neglecting the problem? Diversity and Distributions 24:1883-1889.

Peter BM, Petkova D, Novembre J. 2020. Genetic landscapes reveal how human genetic diversity aligns with geography. Molecular Biology and Evolution 37:943-951.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. Current Biology 15:R159-R160.

R Core Team. 2013. R: A language and environment for statistical computing.

Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, Ripley MB. 2013. Package ‘mass’. Cran R 538.

Roberts A, McMillan L, Wang W, Parker J, Rusyn I, Threadgill D. 2007. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. Bioinformatics 23:i401-i407.

Root M. 2000. How we divide the world. Philosophy of Science 67:S628-S639.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1:e70.

Schwender H. 2012. Imputing missing genotypes with weighted k nearest neighbors. Journal of Toxicology and Environmental Health, Part A 75:438-446.

Serre D, Pääbo S. 2004. Evidence for gradients of human genetic diversity within and among continents. Genome research 14:1679-1685.

Sloan CD, Duell EJ, Shi X, Irwin R, Andrew AS, Williams SM, Moore JH. 2009. Ecogeographic genetic epidemiology. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 33:281-289.

Sugiyama M. 2007. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. Journal of machine Learning research 8:1027-1061.

Sugiyama M editor. Proceedings of the 23rd international conference on Machine learning. 2006.

Vapnik V. 1998. The support vector method of function estimation. In. Nonlinear Modeling: Springer. p. 55-85.

Wang C-C, Yeh H-Y, Popov AN, Zhang H-Q, Matsumura H, Sirak K, Cheronet O, Kovalev A, Rohland N, Kim AM. 2021. Genomic insights into the formation of human populations in East Asia. Nature:1-10.

Weston J, Schölkopf B, Eskin E, Leslie C, Noble WS. 2003. Dealing with large diagonals in kernel matrices. Annals of the Institute of Statistical Mathematics 55:391-408.

Wilkinson S, Haley C, Alderson L, Wiener P. 2011. An empirical assessment of individual-based population genetic statistical techniques: application to British pig breeds. Heredity 106:261-269.

Xiang L, Stillwell J, Burns L, Heppenstall A. 2020. Measuring and Assessing Regional Education Inequalities in China under Changing Policy Regimes. Applied Spatial Analysis and Policy 13:91-112.

Yang MA, Fan X, Sun B, Chen C, Lang J, Ko Y-C, Tsang C-h, Chiu H, Wang T, Bao Q. 2020. Ancient DNA indicates human population shifts and admixture in northern and southern China. Science 369:282-288.

Yang W-Y, Novembre J, Eskin E, Halperin E. 2012. A model-based approach for analysis of spatial structure in genetic data. Nature genetics 44:725.

Yang W-Y, Platt A, Chiang CW-K, Eskin E, Novembre J, Pasaniuc B. 2014. Spatial localization of recent ancestors for admixed individuals. G3: Genes, Genomes, Genetics 4:2505-2518.

Zelnik-Manor L, Perona P. 2004. Self-tuning spectral clustering. Advances in neural information processing systems 17:1601-1608.

Fig.1

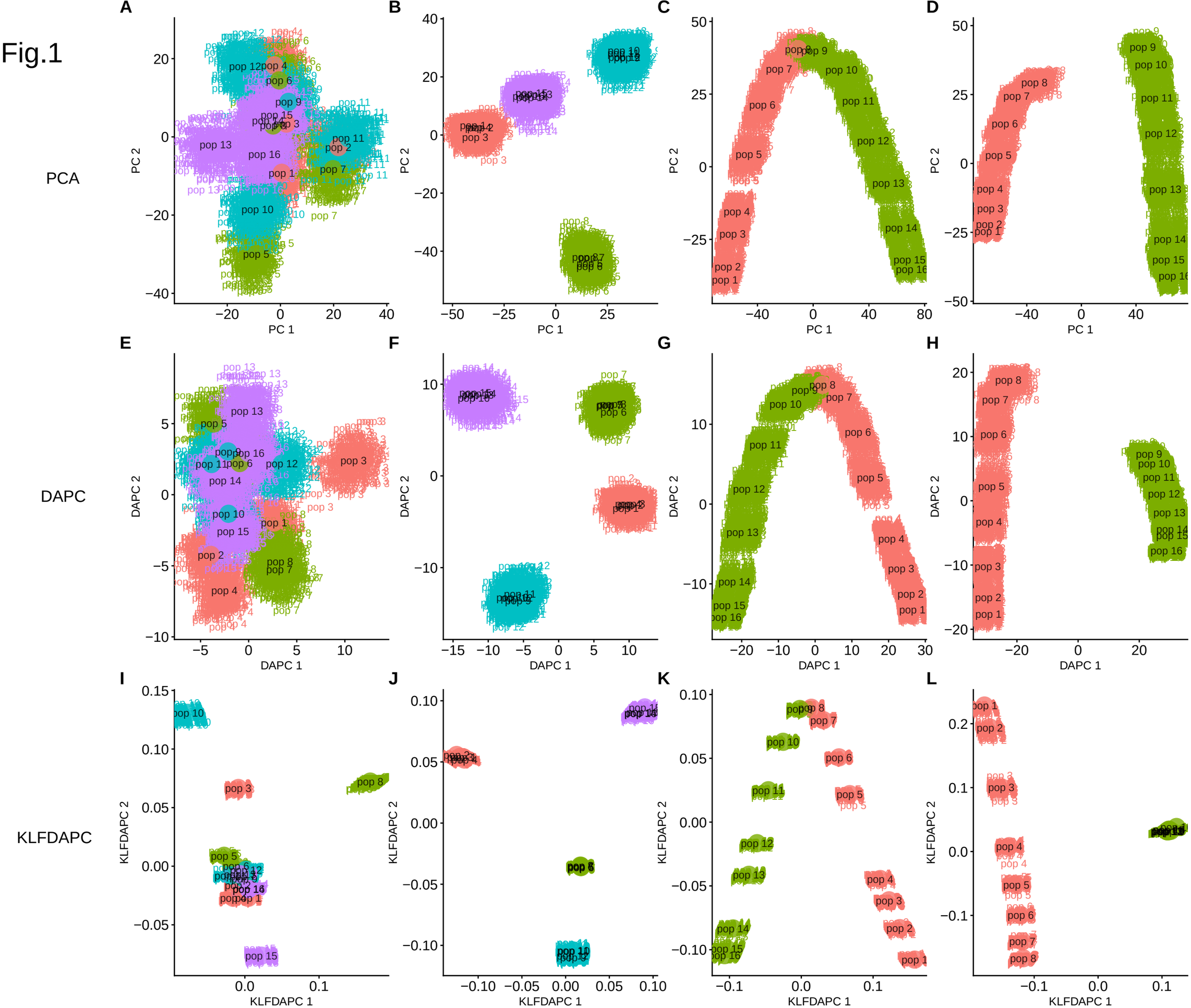
A
B
C
D
E
F
G
H
I
J
K
L
PCA
DAPC
KLFDAPC
PC 1
PC 2
DAPC 1
DAPC 2
KLFDAPC 1
KLFDAPC 2

Fig. 2

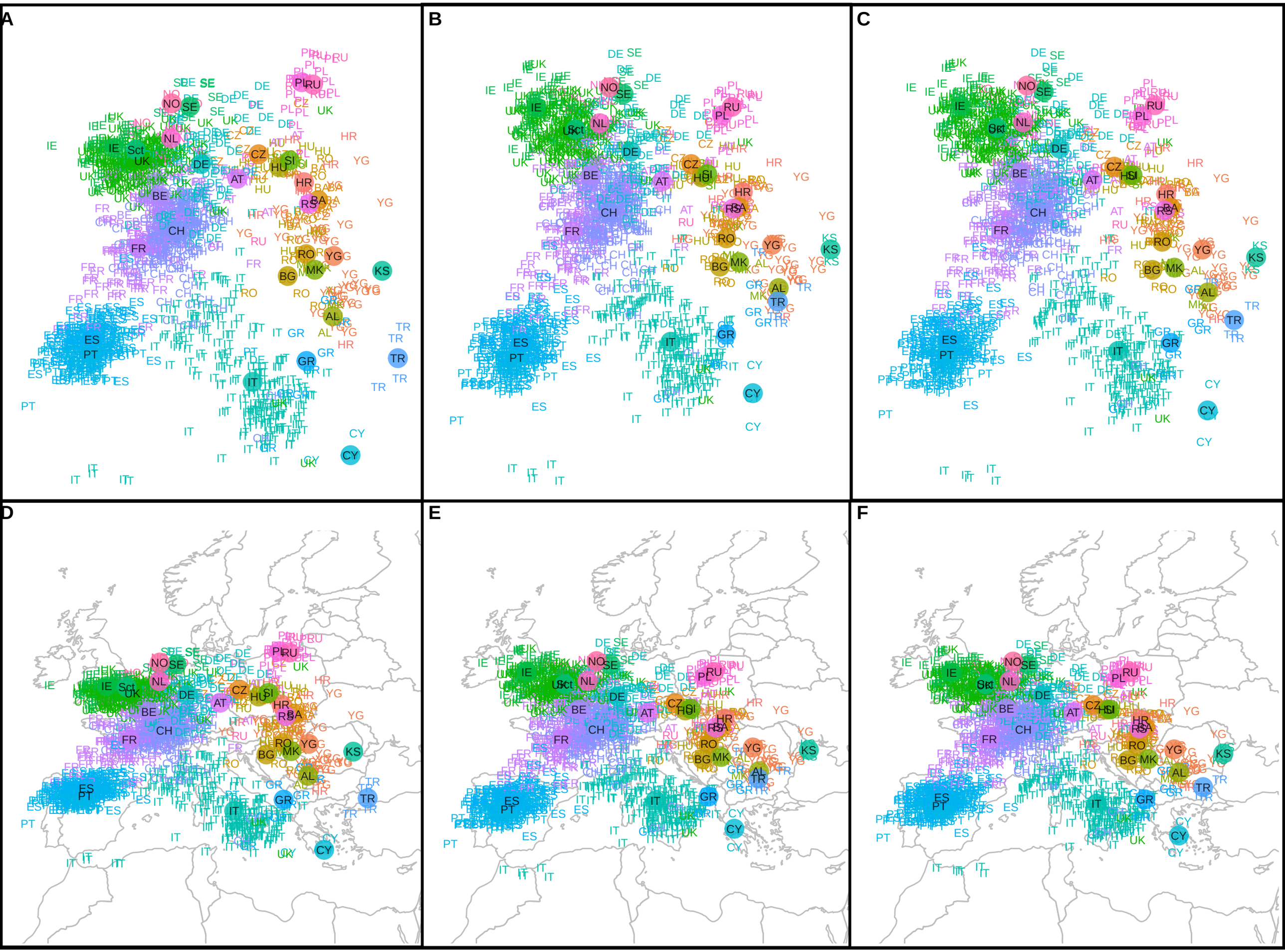
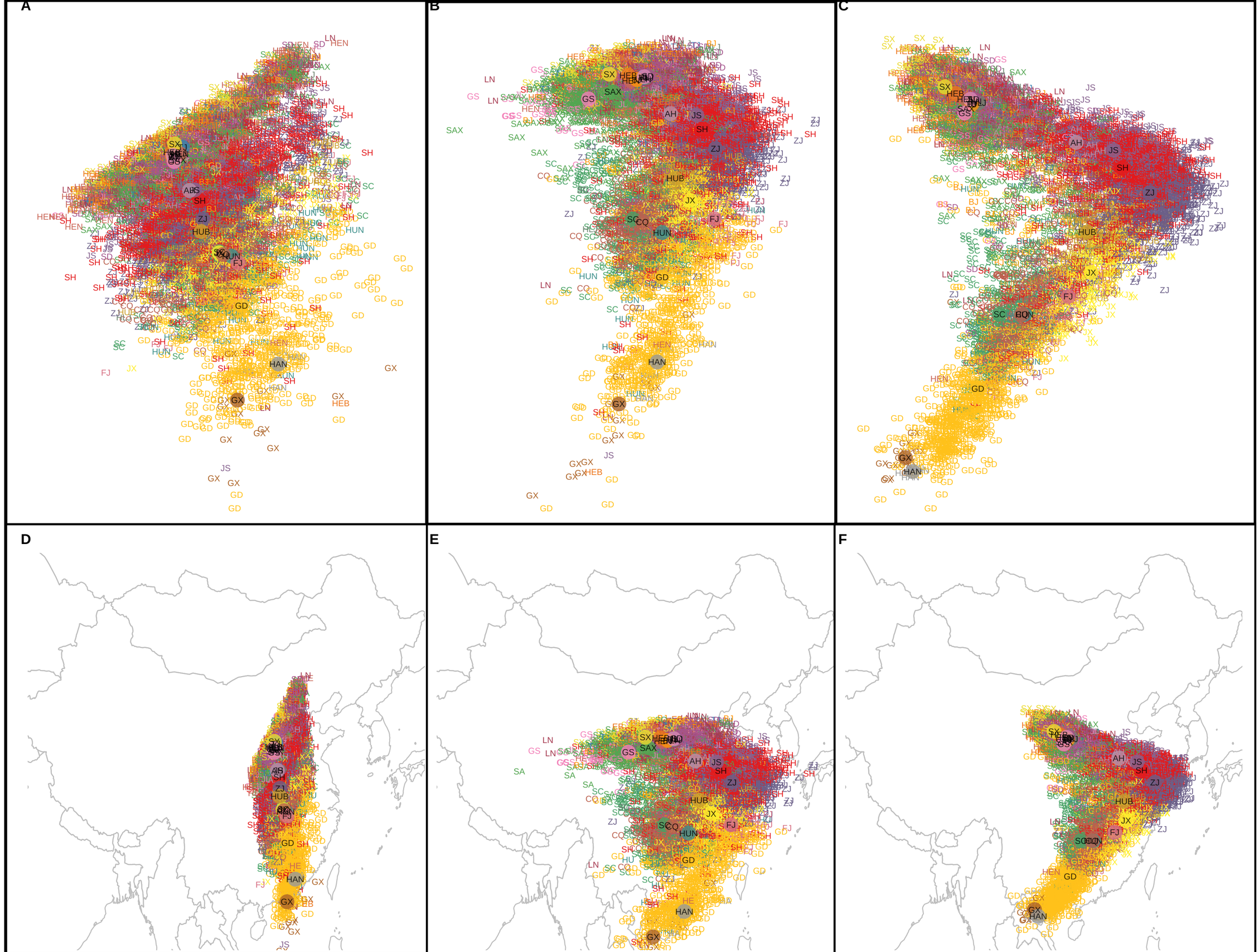
Fig. 3