The leading *k* – 1 PC axes for *k* effective populations are the only dimensions suitable for DAPC of genotype data

Joshua A. Thia [1, *]

[1] Bio21 Institute, School of BioSciences, The University of Melbourne, AU

* Corresponding author: joshua.thia@unimelb.edu.au; josh.thia@live.com

Running title: DAPC parameterisation with *k* – 1 PC axes

Keywords: coalescent simulations; discriminant analysis; multivariate statistics; population genetic methods; principal component analysis.

Abstract

Multivariate methods are incredibly beneficial for population genetic analyses where the number of measured variables (genetic loci) can easily exceed the number of sampled individuals. Discriminant analysis of principal components (DAPC) has become a popular method for visualising population structure in genotype datasets due to its simplicity, its computational speed, and its freedom from demographic assumptions. Despite the popularity of DAPC in population genetic studies, there has been little discussion on best practise and parameterisation. Unappreciated, perhaps, is the fact that unlike principal component analysis (PCA), which is a hypothesis free method, discriminant analysis (DA) is a hypothesis driven method. That is, when performing a DA, a researcher is making an explicit hypothesis about how variation in a set of predictor variables is organised among pre-defined groups in a sample set. Parameter choice is critical to ensure the results produced by a DA are biologically meaningful. In a DAPC, one of the most important parameter choices is the number of PC axes, p_{axes}, to use as predictors in a DA of among-population differences. Yet there are no clear guidelines on how researchers should choose p_{axes}. In this work, I propose that the maximum value of p_{axes} is a deterministic feature of a genotype dataset based on population genetic theory. For *k* effective populations, only the first *k* – 1 PC axes are expected to be biologically informative and capture population structure. DA fit using more than the first *k* – 1 PC axes are over-fit and may discriminate groups using biologically uninformative predictors. Using samples drawn from simulated metapopulations, I show that DAPC parameterised using no more than the leading *k* – 1 PC axes are: (1) more parsimonious; (2) capture the maximal amount of among-population variation using biologically relevant predictors; (3) are less sensitive to unintended interpretations of population structure; and (4) are more generally applicable to independent sample sets. I provide practical recommendations for selecting an appropriate p_{axes} for DAPC relative to the putative number of effective populations and a researcher's hypothesis about how genetic variation is organised.

Introduction

The biological world is beautifully complex, characterised by variation in multiple dimensions. Multivariate statistics play a pivotal role in helping us make sense of this multi-dimensionality and developing a deeper appreciation of biology. Describing population genetic patterns, for example, becomes increasingly difficult with many sampled individuals, genetic markers, and populations. However, ordination methods can summarise variation across multiple loci to create new synthetic axes and reduce dimensionality. Such new axes of variation facilitate the visualisation of genetic relationships among individuals and populations, aiding the interpretation of population structure.

Principal component analysis (PCA) is an incredibly useful multivariate method for summarising genetic differences among individuals. PCA reveals structure in a dataset by capturing the major axes of covariance among a set of *p* variables (Rencher, 2002b). Putative groupings can be inferred from the clustering of samples in discrete regions of PC space. Because PCA does not make any explicit assumptions about the organisation of variation in the *p* variables, PCA is a *hypothesis free* method, providing information about the innate structure in a sample (Rencher, 2002b). All PC axes are orthogonal and are constrained to capture decreasing amounts of variance from the first to last axis (Rencher, 2002b). Hence, in a structured set of variables, dimension reduction can be achieved by selecting some number of the leading PC axes that describe the most variation (Rencher, 2002b). The dimensionality reduction achieved by a PCA is hugely beneficial in population genomic studies, summarising the covariances among thousands, maybe millions, of loci (Patterson, Price, & Reich, 2006). Determining which PC axes capture biologically informative variation is a non-trivial task (Cattell, 1966; Jackson, 1993). However, for a genotype dataset comprising *k* effective populations, prior theoretical work indicates that only the first $k - 1$ PC axes capture population structure (Patterson et al., 2006).

Discriminant analysis (DA) is another multivariate method that has become popular in population genetic studies through the discriminant analysis of principal components (DAPC) approach (Jombart, Devillard, & Balloux, 2010) (Figure 1). DA takes a set of measured *p* predictor variables and identifies linear combinations of those variables that maximise the discrimination of individuals from pre-defined groups. Because DA explicitly searches for differences among groups in their multivariate means, DA is a *hypothesis driven* method (Rencher, 2002a). Performing a DA directly on a genotype dataset of many loci is undesirable because the number of *p* variables (genetic loci) often greatly exceed the number of *n* sampled individuals, and there may be extensive correlations among loci (Jombart et al., 2010). The DAPC approach circumvents this issue by first performing a PCA to reduce dimensionality and remove correlation, with the first p_{axes} number of PC axes used as predictors in to discriminate among k_{DA} groups in a DA. The value of k_{DA} chosen by a researcher represents their hypothesis about how genetic variation is organised in their dataset. For a DAPC, k_{DA} is chosen in one of two ways: (1) $k_{DA} = k_{prior}$, where k_{prior} is an *a prior* expectation about the number of populations, for example, the number of sampling locations; or (2) $k_{DA} = k_{infer}$, where k_{infer} is an inference of the number of effective *k* populations, obtained from genotype data.

Despite the frequent use of DAPC for summarising population genetic structure, there has been very little discussion of best practise associated with this approach. Such matters were recently addressed in a combined meta-analysis and simulation study by Miller et al. (2020). In their work, Miller et al. (2020) highlight that many studies fail to report the relevant parameters required to replicate their DAPC results. They also provided some practical suggestions on how DAPC of genotype data can be made more transparent by reporting of how groups are pre-defined, clearly specifying the goals of a DAPC, and stating the number of PC axes used in the analysis.

In this piece, I address the consideration of the number of PC axes, p_{axes}, used to parameterise a DAPC. Typically, in population genetic studies, the chosen value of p_{axes} is very large, with the intention of including a substantial proportion of the total genotypic variance in the DA model. Yet based on the expectation of $k - 1$ biologically informative PC axes for k effective populations (Patterson et al., 2006), the use of $p_{axes} > k - 1$ is over-fit. Here, I propose that the $k - 1$ limit of biologically informative PC axes sets a deterministic value on the maximum p_{axes} that should be used in a DAPC of genotype data. I illustrate this point using a series of analyses conducted on samples drawn from simulated metapopulations that vary in their migration rates and genetic differentiation. I demonstrate that the discriminative power in a DA model comes exclusively from the leading $k - 1$ PC axes, and that inclusion of many more than $k - 1$ PC axes inflates the perceived differences among k_{DA} hypothesised groups. These results indicate that parameterisation of a DAPC would benefit from attempts to choose an appropriate p_{axes} relative to the effective number of populations, k, and a researcher's hypothesis, k_{DA}.

(Note, for reference, relevant mathematical notations are detailed in Table 1.)

Simulated population genetic datasets

I simulated metapopulations comprising five populations connected by gene flow using *fastsimcoal2* v2.7 (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013; Excoffier et al., 2021). Each population was comprised of 1,000 haploid genomes (500 diploids). Each genome was 20 Mb in length with a recombination rate between adjacent bases of 1e−8 and a mutation rate of 2e−9. Populations underwent a series of hierarchical divergence events, with each divergence event 10,000 generations from the next (Figure S1). I modelled three migration scenarios with different rates of gene flow, with M as the overall rate of gene flow into a focal population, and m as the individual contribution of each source population into a focal population: (a) $M = 0.000$, $m = 0.000$, divergence with no gene flow, $F_{ST} = 0.99$; (b) $M = 0.004$, $m = 0.001$, divergence with gene flow, $F_{ST} = 0.09$; and (c) $M = 0.040$, $m = 0.010$, an effectively panmictic (homogeneous) metapopulation, $F_{ST} = 0.01$ (see Figures S2 and S3 for simulation summaries). These scenarios reflect different contexts where the number of perceived sampled populations may or may not be equal to the number of k effective populations. In the $M = 0.000$ scenario, there is strong population structure, and a clear $k = 5$. In the $M = 0.004$ scenario, population structure is weaker, but there is still a discernible $k = 5$. In the $M = 0.040$ scenario, there is no population structure, and $k = 1$.

For each migration scenario, I created two distinct simulated datasets. The first involved a single simulation at each migration rate where all 500 diploid contemporaneous individuals per population were sampled at the end of the simulation: herein, the "singular simulation dataset". The purpose of this first dataset was to illustrate variation and relationships among independent sample sets from the same simulated metapopulation. Diploid individuals were split into 20 sample sets, each containing 25 individuals per population. In the second dataset, I performed 30 independent simulations at each migration rate and sampled 25 diploid contemporaneous individuals per population at the end of the simulation: herein, the "replicate simulation dataset". The purpose of this second dataset was to illustrate generalities across independent metapopulations that had undergone the same migration scenario.

Simulation results were imported into R v4.1.3 (R Core Team, 2022). In each simulation in the *replicate simulation dataset*, I filtered SNP loci to a sample size of 1,000 random SNPs distributed across the genome. First, the genome was divided into 5 Kb windows, from which one random SNP was sampled. Second, these remaining SNPs were randomly sampled to retain 1,000 loci. For the *singular simulation dataset*, a preliminary filtering was performed to remove all loci that were not polymorphic across all discrete sample

sets. Following this, random sampling per genomic window and reduction to 1,000 random loci was performed. Details of all downstream analyses can be found in the Supplementary Information.

Population structure captured by principal component analysis

Determining which PC axes, if any, capture significant covariance among measured variables presents a major challenge for the use of these dimensions in predictive analyses (Peres-Neto, Jackson, & Somers, 2005). Patterson et al. (2006) demonstrated a clear expectation that for k effective populations there are $k - 1$ PC axes that capture variation among populations, with all remaining axes characterising variation within populations. The inference of k, k_{infer}, is obtained from the genotype data. In this section, I will highlight simple ways that k_{infer} can be obtained visually and explain why we should only focus on the leading $k - 1$ PC axes of population structure. Although various methods exist to statistically test the number of biologically informative PC axes (Jackson, 1993; Peres-Neto et al., 2005) and to obtain k_{infer} from genotype data (Patterson et al., 2006), I will not address the performance of these methods here. I will, however, discuss the use of *K*-means clustering to obtain k_{infer}, as it is commonly used in conjunction with DAPC.

One simple way to visually obtain k_{infer} is through examining PC scree plots. Scree plots illustrate the decline in explained variance (or eigenvalues) associated with each sequential PC axis. When population structure exists, there is typically an inflection point at the kth PC axis that creates an "elbow shaped" pattern in the explained variances (Abegaz et al., 2019; Cattell, 1966), as illustrated in Figure 2a. We can see that in the $M = 0.000$ and $M = 0.004$ scenario, the explained variance rapidly decreases from PC axes 1 to 4 ($k - 1$), and then decreases more incrementally beyond PC axis 5 (k). The scree is also steeper between PC axes 1 to 4 in the $M = 0.000$ scenario because more variation is among populations (F_{ST} approaching 1.00), relative to that in the $M = 0.004$ scenario. Contrastingly, in the $M = 0.040$ scenario, although there were five sampled populations, there is no variation among them ($k = 1$), and the scree slope exhibits a smooth decline.

A second way to visually obtain k_{infer}, and to validate inferences from scree plots, is by examining the scatter of individuals projected into PC space. Each dimension from PC axis 1 to $k - 1$ should capture different aspects of the among-population variation for k effective populations. PC axes $\geq k$, however, are associated with stochastic sampling noise and within-population variation. Therefore, clustering of samples should only be observed on PC axes $\leq k - 1$. These expectations are illustrated in Figure 3. For $M = 0.000$ on PC axes 1 to 4 (Figures 3a and 3b), individuals from each population pile up into very discrete regions of PC space because almost all the variation is among populations. For $M = 0.004$ on PC axes 1 to 4 (Figures 3d and 3e), populations clusters are more dispersed, and their multivariate means are closer in PC space, because genetic differentiation is weaker. For both $M = 0.000$ and 0.004, beyond PC axis 5 (Figures 3c and 3f, respectively), all populations collapse into a single homogeneous data cloud. In contrast, for the $M = 0.040$ scenario, there is no clustering of individuals in PC space on the first four PC axes, and all the variation is within populations because $k = 1$.

K-means clustering offers another way to obtain k_{infer}. The goal of *K*-means clustering is to allocate observations into a specified number of $k_{cluster}$ clusters by minimising the within-group sum of squares (J. A. Hartigan and M. A. Wong, 1979). This method is commonly implemented within a DAPC pipeline, although it is not an essential element of DAPC. Many values of $k_{cluster}$ can be fitted and compared using information criteria, for example, a BIC score. The most likely value is then chosen to represent k_{infer}, which can then be used to inform the hypothesis, k_{DA}, in the DA step. It is important to note that unlike PC scree plots and scatter plots where we are directly examining the raw output of a PCA to obtain k_{infer}, *K*-means clustering is

a method applied to the data that requires parameterisation. In addition to the $k_{cluster}$ parameter, a set of predictor variables needs to be specified, which is typically the same of leading PC axes, p_{axes}, that will be used in DAPC. Figure 4 illustrates the range of *K*-means clustering results using different values of $k_{cluster}$ and p_{axes} as predictors, with cluster fit evaluated using BIC scores (lower is more likely). For the $M = 0.000$ scenario, the BIC scores rapidly decline and plateau at $k_{cluster} = 4$ in a relatively consistent way for all values of p_{axes}. For the $M = 0.004$ scenario, the BIC scores exhibit an inflection at $k_{cluster} = 4$, although the shape of the BIC curves differ depending on p_{axes}: whereas there is a plateauing pattern for $p_{axes} = 10$ and 20, there is an uptick pattern for $p_{axes} = 40$ and 80. For $M = 0.040$, there are also two different BIC curves: for $p_{axes} = 10$ and 20, BIC exhibits a steady declining pattern (inconsistent with $k = 1$), whereas for $p_{axes} = 40$ and 80, BIC exhibits a steading increasing pattern (consistent with $k = 1$). Notably, whilst *K*-means clustering should generally reflect the underlying *k*, different parameterisations can lead to different patterns in the test statistic curves. If *K*-means is used to obtain k_{infer}, this estimate should also be compared to expectations derived from PC scree plots and scatter plots to assess consistency across approaches.

Aside from the statistical principle that only the first $k - 1$ PC axes capture population structure, these leading PC axes are also the only dimensions that are replicable. Replicability of PC axes in this context refers to recoverability of the same (or similar) eigenvectors across independent samples. Eigenvectors describe the relative contribution of each locus to each PC axis (the loadings). Loci contributing most to population structure should be consistent in the magnitude of their loadings. Where population structure exists, eigenvectors of the leading PC axes $\leq k - 1$ are correlated across independent sample sets, whereas eigenvectors of PC axes $\geq k$ are not correlated across independent sample sets, as depicted in Figure 2c. However, the recoverability of eigenvectors depends on the magnitude of genetic differentiation. As genetic differentiation increases: (1) the intra-locus allelic correlation between individuals from the same population increases (Wright, 1949); (2) sampling bias is less likely to affect estimates of the inter-locus allelic covariances across all individuals; and (3) eigenvectors for the first $k - 1$ PC axes will be more similar across independent sample sets. Compare, for example, the correlations between sample pair eigenvectors for the first four PC axes where $M = 0.000$ versus $M = 0.004$ versus $M = 0.040$, which show a decline with decreasing genetic differentiation. Moreover, in the $M = 0.040$ scenario, none of the eigenvectors have high correlations across independent sample sets because there is no population structure and no linear combinations of alleles that are repeatable. We should be concerned with this repeatability if we wish to use these PC axes as predictive variables, as discussed in the next section.

In summary, the analyses presented in this section illustrate how *k* can be inferred from a PCA of genotype data. Obtaining an inference of *k*, k_{infer}, is important because only the leading $k - 1$ PC axes are biologically informative and exhibit repeatable linear combinations of loci across independent sample sets. PC axes $\leq k - 1$ capture variation that is among populations, whereas PC axes $\geq k$ are associated with variation that is within populations. Obtaining k_{infer} is also useful because it can inform whether some initial *a priori* expectation of the number of populations, k_{prior}, is valid.

Considerations for discriminant analysis of principal components

In this section, I will address the appropriate parameterisation of a DAPC. I will first discuss common practises in DAPC paramterisation and their (mis)alignment with the $k - 1$ limit of biologically informative PC axes. I will then provide a demonstration of how different parameterisations of a DAPC influence the solutions derived from its DA step. These analyses will highlight that only those PC axes $\leq k - 1$ are suitable

for DAPC. Inappropriate parameterisation of DAPC reduces the biological relevancy of the DA solution and can give false impressions of population structure.

That we only expect the leading $k - 1$ PC axes are biologically informative implies that the maximum value of p_{axes} should be $k - 1$. However, DAPC parameterisation with the $k - 1$ limit in mind is not common place. Instead, typical DAPC parameterisation often involves choosing a value of p_{axes} that captures a certain proportion of the total genotypic variance. Another common approach is to use the built-in functions from the *adegenet* package, XVALDAPC and OPTIM.A.SCORE (Jombart, 2008; Jombart & Ahmed, 2011), to reduce and optimise the number of PC axes predictors, p'_{axes}. The XVALDAPC function performs cross-validation for different numbers of p_{axes}, inferring the optimal p'_{axes} as the one that produces the lowest mean squared error term. The OPTIM.A.SCORE function performs a reassignment analysis using true and randomised cluster identities to calculate an *a*-score for different numbers of p_{axes}, inferring the optimal p'_{axes} as the one that produces the largest *a*-score.

Selecting p_{axes} to satisfy a proportional variance criterion is problematic. Unless there is very strong differentiation among populations, the value of p_{axes} required to capture a specific amount of genotypic variance will likely be much greater than $k - 1$. Additionally, the built-in optimisation functions have variable performance that may depend on the magnitude of genetic differentiation and the initial value of p_{axes}, as illustrated in Figure 4. In my simulations, the XVALDAPC function consistently returned a p'_{axes} that was greater than the $k - 1 = 4$ expectation for $M = 0.000$ and $M = 0.004$ scenarios, and greater than the $k - 1 = 0$ expectation for $M = 0.040$, for all initiating values of p_{axes}. The OPTIM.A.SCORE function tended to be more conservative, but increasing migration rates and larger values of p_{axes} led to inflated estimates of p'_{axes}.

Because population structure is only captured on the leading $k - 1$ PC axes, the addition of axes beyond $k - 1$ does not increase the variation among populations explained by a DA model. Figure 5 depicts the mean variation among populations when using different values of p_{axes} to parameterise a DAPC with the hypothesis, $k_{DA} = 5$. For $M = 0.000$, irrespective of the value of p_{axes}, the amount of variation among populations is always >95% because the first four PC axes explain virtually all the genotypic variance. However, in the $M = 0.004$ scenario, increasing p_{axes} reduces the variation among populations. Finally, for $M = 0.040$, because there is no population structure, none of the PC axes capture among-population variation, and there is essentially a flat line across all values of p_{axes}. When population structure exists, limiting p_{axes} to the leading $k - 1$ PC axes for DAPC parameterisation is more parsimonious, uses only biologically relevant predictors, and maximises the total variation among populations present in the input predictors.

Although biologically informative PC axes should contribute most to the DA solution, as p_{axes} increases beyond $k - 1$, the DA solution can be influenced by biologically uninformative predictors. Consider the contributions of PC axes in highly over-fit DA models using $p_{axes} = 80$ PC axes with $k_{DA} = 5$. Figure 6 depicts the average absolute loadings of the first 12 PC axes onto the discriminant functions that describe the four-dimensional LD space. For the $M = 0.000$ and $M = 0.004$, PC axes 1 to 4 ($\leq k - 1$) have consistently larger loadings relative to PC axes 5 and beyond ($\geq k$). Notably, for $M = 0.000$, each sequential LD axis largely correspond to each PC axis in descending order from 1 to 4, as evidenced by the large absolute loadings. For $M = 0.000$, there is minimal within-population variation, so PC axes ≥ 5 have negligible contributions to the LD axes, and DA effectively return the PCA results. For the $M = 0.004$ scenario, the first four PC axes combine in different ways to generate the LD axes. Unlike $M = 0.000$, DA does not simply return PCA results for $M = 0.004$, and the uninformative PC axes ≥ 5 are having a larger influence on the final solutions. In the $M = 0.040$ scenario, $k - 1 = 0$, the LD loadings are equivalent for all PC axes. For $M = 0.040$, all the variation is within populations, but the DAs derive solutions despite none of the predictor PC axes capturing

population structure. These data demonstrate that when genetic differentiation is greater, a DA is less likely to be influenced by the inclusion of many biologically uninformative PC axes. In contrast, when genetic differentiation is weaker, inclusion of many biologically uninformative PC axes > k – 1 is more likely to influence a DA. Furthermore, because of the hypothesis driven nature of a DA, a solution will be derived whether it is biologically meaningful or not.

By inspecting scatter plots of individuals projected into LD space, we can see how genetic differentiation and the choice of k_{DA} and p_{axes} interact in a DAPC (Figure 7). For all migration scenarios, increasing p_{axes} increases the separation of populations in LD space when the hypothesis is k_{DA} = 5; note the increasing scale of the LD axes and tighter clustering. However, with respect to the general patterns inferred from the LD space projection, the effect of increasing p_{axes} > k – 1 is more dramatic when genetic differentiation is weak. For M = 0.000, regardless of whether p_{axes} = 4, 40 or 80, we essentially recover the same projection of populations in LD space (compare Figures 7a, 7b, and 7c). For M = 0.004, the same general pattern emerges regardless of the value of p_{axes}, but the perceived magnitude of this separation depends on p_{axes} (compare Figures 7d, 7e, and 7f). For M = 0.040, despite k = 1 and a lack of population structure, increasing p_{axes} gives the impression that populations can be reliably discriminated into discrete clusters. When using p_{axes} = 4 (Figure 7g), we retain a more realistic projection with negligible separation among populations in LD space. However, with p_{axes} = 40 and 80 (Figures 7h and 7i, respectively), there is increasing separation of populations in LD space, which is driven by the inclusion of many biologically uninformative PC axes into the DA solution. Hence, the degree to which p_{axes} affects our interpretation of population structure from a DAPC may depend on: (1) the magnitude of genetic differentiation; and (2) the divergence of p_{axes} from k.

Because DA is often used for group assignment problems, I will use group assignment to further explore how different combinations of PC axes contribute (or do not contribute) to discrimination among populations. Our expectation is that when population structure exists, DA models including the first k – 1 PC axes will have better ability to assign individuals to their correct population, relative to DA models without these biologically informative PC axes. Figure 8 illustrates assignment analyses where a single *training* sample set was used to build a DA model with the hypothesis of k_{DA} = 5, and this model was then used to predict the populations of individuals in 19 *testing* sample sets. Different combinations of PC axes were used that either included the first four PC axes (1, 1–2, 1–3, 1–4, 1–40, and 1–80) or excluded them (5–40 and 5–80). Clearly, in the M = 0.000 and M = 0.004 scenarios (Figures 8a and 8b, respectively), the power to assign individuals to their correct population is being completely driven by the leading four PC axes. Exclusion of the first four PC axes leads to poor correct assignment rates in these scenarios where population structure exists. For the M = 0.040 scenario, there is no power to assign individuals, irrespective of the input p_{axes}.

The results from the assignment analyses (Figure 8) can be further interpreted by considering the recoverability of eigenvectors that were discussed above (Figure 2c). The linear combinations of loci that are described by eigenvectors are more repeatable across independent sample sets when there is greater genetic differentiation, but only for those k – 1 PC axes describing population structure (Figure 2c). In the context of a group assignment problem, whereas the first k – 1 PC axes provide repeatably useful predictors to assign populations, PC axes ≥ k are stochastic across different sample sets and are uninformative predictors for population assignment. The M = 0.000 scenario provides an extreme case where k = 1, there are no biologically informative PC axes, none of the eigenvectors are repeatable, and assignment of individuals to the five populations is not possible.

In summary, the analyses presented in this section demonstrate that only the leading k – 1 PC axes are suitable for DAPC of genotype data. The hypothesis driven nature of a DA requires appropriate

296 parameterisation to ensure biologically meaningful results. Addition of PC axes greater than $k - 1$ does not
297 add informative predictors to discriminate among k_{DA} hypothesised groups in a DA. When genetic
298 differentiation is weak, the inclusion of many uninformative PC axes produces an over-fit model that
299 artificially inflates the perceived separation among groups in LD space. DAPC parameterised using only the
300 leading $k - 1$ PC axes are maximised for the among-population variation in the predictor variables, are not
301 influenced by the idiosyncrasies of uninformative predictors, and are generalisable to independent sample
302 sets.

303

304 Concluding remarks

305 As a hypothesis free method, PCA is useful for visualising the inherent structure present in a dataset.
306 Contrastingly, DA is a hypothesis driven method and presents a visualisation that maximises the differences
307 among pre-defined groups. The combination of these methods into a DAPC of genotype data has become
308 very popular. DAPC provides a fast and simple method to summarise population structure and to perform
309 population assignment. The ease at which DAPC can be applied to genotype data in R, and its lack of
310 demographic assumptions, make it a wonderfully flexible analysis. Yet the lack of clear best practise
311 guidelines has made the implementation, interpretation, and reporting of DAPC results inconsistent (Miller
312 et al., 2020).

313 Inclusion of many biologically uninformative PC axes artificially inflates the separation among the
314 hypothesised k_{DA} groups in LD space. This is problematic because LD space projections are often a focal
315 point of interpreting the results of a DAPC. Yet if separation among groups is artificially inflated, there is the
316 risk that the perceived population structure is greater than that present in the genotype dataset. Only the
317 leading $k - 1$ PC axes for k effective populations are suitable for DAPC as they are the only dimensions
318 capturing population structure. Whereas other methods implementing PCA perform a selection step to
319 limit analyses to biologically informative PC axes (for example: Conomos, Reiner, Weir, & Thornton, 2016;
320 Luu, Bazin, & Blum, 2017; Meisner & Albrechtsen, 2018), such considerations have never been discussed in
321 the context of performing a DAPC, to the best of my knowledge. This is surprising given prior works
322 describing the link between PCA and the genetic relationships among individuals and populations (McVean,
323 2009; Patterson et al., 2006; Peter, 2022). There are many ways to estimate the number of biologically
324 informative PC axes (Cattell, 1966; Jackson, 1993; Peres-Neto et al., 2005). For genotype data, Patterson et
325 al. (2006) provided a test based on expectations under a Tracy-Widom distribution to statistically infer k.
326 Notwithstanding, careful examination of PC scree plots and scatter plots provide a simple visual way to
327 infer k, and this inference can be corroborated with estimates from K-means clustering.

328 Practically, if a researcher has an *a prior* expectation of the number of populations, k_{prior}, they should also
329 infer k from their genotype data, k_{infer}, to validate their prior expectations. When k_{prior} and k_{infer} are
330 mismatched, a researcher must decide whether they are still interested in testing the hypothesis of k_{DA} =
331 k_{prior} in their DAPC, or whether $k_{DA} = k_{infer}$ is more appropriate. A researcher might have specific questions
332 about the relationships among sampling locations, for example, that might motivate them to test k_{prior},
333 even when $k_{prior} \neq k_{infer}$. Parameterising a DAPC with $k_{DA} = k_{prior}$ when $k_{prior} \neq k_{infer}$ should entail a choice of
334 p_{axes} that: (1) is does not exceed the biologically informative PC axes for k; and (2) does not over-fit with
335 respect to k_{DA}. Under more complex demographic scenarios, it may be challenging to estimate k_{infer} when
336 populations do not form easily discernible groups (see for illustration, Figure S5). Nonetheless, the
337 expectations of a $k - 1$ limit on the number of biologically informative PC axes will still hold, and a

researcher should do their best to make an appropriate judgement of k_{infer} to inform their choice of p_{axes} for DAPC.

Suggestions for future population genetic studies include:

1. Clearly state the rationale for the hypothesised number of groups in a DAPC, k_{DA}, and indicate whether k_{DA} reflects *a priori* expectations, k_{prior}, or the inferred number of effective populations, k_{infer}.
2. Verify that results from *K*-means clustering are consistent with expectations from PC scree plots and scatter plots when obtaining k_{infer}.
3. Where there is no *a priori* expectation, and $k_{DA} = k_{infer}$ is used in a DAPC, limit discrimination to $p_{axes} = k_{infer} – 1$ PC axes.
4. Where there is an *a priori* expectation, and $k_{DA} = k_{prior}$ is used in a DAPC, estimate k_{infer} to determine whether: (a) $k_{prior} = k_{infer}$, (b) $k_{prior} < k_{infer}$, or (c) $k_{prior} > k_{infer}$.
5. When $k_{DA} = k_{prior}$ is used in a DAPC, and $k_{prior} = k_{infer}$, limit discrimination to $p_{axes} = k_{prior} – 1$ PC axes.
6. When $k_{DA} = k_{prior}$ is used in a DAPC, but $k_{prior} < k_{infer}$, limit discrimination to $p_{axes} = k_{prior} – 1$ PC axes. This is a conservative choice because there are more effective genetic groups than the *a priori* expectation.
7. When $k_{DA} = k_{prior}$ is used in a DAPC, but $k_{prior} > k_{infer}$, limit discrimination to $p_{axes} = k_{infer} – 1$ PC axes. This choice ensures that only biologically informative predictors are used.
8. Check that k_{infer} is not artificially inflated, for example, by the presence of highly correlated genotypes in sex chromosomes or inversions, or by inclusion of closely related kin. Tightly linked loci should be removed or recoded as a single “super locus”, and closely related kin should be excluded.
9. When $k_{infer} = 2$, there is just a single dimension that is biologically informative: the first PC axis. A DAPC can still be fit with a single predictor variable, although it defeats the purpose of using a multivariate analysis and is equivalent to a univariate analysis.
10. Cross-validation can be used to assess the DA fit and the relationships among the hypothesised k_{DA} groups. Ideally, cross-validation would use an independent sample set to test a DA fitted to a training sample set. Alternatively, a leave-one-out approach can also be used when a testing sample set is unavailable.

It is perhaps prudent to reflect on whether a DAPC is necessary for the specific goals of a study. If the objective is to simply visualise population structure, a PCA alone is perfectly sufficient and does not require allocation of group membership. However, when there are many putative populations, visualising just a few PC axes might be inadequate to fully appreciate the complexity of population structure. In such a case, DAPC may be helpful for summarising variation across many relevant PC axes and to further reduce dimensionality. The real value of a DAPC comes when we take advantage of the predictive model produced by the DA. These predictive models can be used in assignment problems to identify source populations of recent migrants or to infer the parental populations in admixture events. The individual contributions of loci to variation among populations could also be exploited to identify candidates for divergent selection. Irrespective of the goal, limiting p_{axes} to the leading $k – 1$ PC axes will produce a DA fit that is more parsimonious and captures maximal among-population variation from biologically relevant predictor variables. An appropriately parameterised DA is less likely to suffer from unintended interpretations of population structure and is more generalisable to individuals from independent samples sets.

Acknowledgements

I thank Cynthia Riginos and Tom Schmidt for providing feedback on this manuscript. This piece benefitted from the thoughtful critiques of two anonymous reviewers, to whom I am grateful.

Conflicts of interest

I have no conflicts of interest associated with this work.

Data accessibility statement

The data and scripts associated with this manuscript have been deposited into *Dryad* (Thia, 2022).

R packages

adegent (Jombart, 2008; Jombart & Ahmed, 2011)

data.table (Dowle & Srinivasan, 2019)

doParallel (Microsoft Corporatin & Steve Weston, 2019)

extrafont (Chang, 2014)

genomalicious (Thia & Riginos, 2019)

ggpubr (Kassambara, 2020)

ggtree (Yu, Smith, Zhu, Guan, & Lam, 2017)

MASS (Venables & Ripley, 2002)

tidyverse (Wickham et al., 2019)

References

Abegaz, F., Chaichoompu, K., Génin, E., Fardo, D. W., König, I. R., John, J. M. M., & Van Steen, K. (2019). Principals about principal components in statistical genetics. *Briefings in Bioinformatics*, *20*(6), 2200–2216. doi: 10.1093/bib/bby081

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Chang, W. (2014). *extrafont: Tools for using fonts*.

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free Estimation of Recent Genetic Relatedness. *American Journal of Human Genetics*, *98*(1), 127–148. doi: 10.1016/j.ajhg.2015.11.022

Dowle, M., & Srinivasan, A. (2019). *data.table: Extension of "data.frame."* Retrieved from https://cran.r-project.org/package=data.table

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905.

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). *fastsimcoal2*: demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*(24), 4882–4885.

J. A. Hartigan and M. A. Wong. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, *74*(8), 2204–2214. doi: 10.2307/1939574

Jombart, T. (2008). *adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Jombart, T., & Ahmed, I. (2011). *adegenet* 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071.

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*(1), 1–15.

Kassambara, A. (2020). *ggpubr: "ggplot2" Based Publication Ready Plots*.

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. doi: 10.1111/1755-0998.12592

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*(10). doi: 10.1371/journal.pgen.1000686

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731.

Microsoft Corporatin & Steve Weston. (2019). *doParallel: Foreach Parallel Adaptor for the "parallel" Package*.

Miller, J. M., Cullingham, C. I., & Peery, R. M. (2020). The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity*, *125*(5), 269–280. doi: 10.1038/s41437-020-0348-2

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093. doi: 10.1371/journal.pgen.0020190

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, *49*(4), 974–997. doi: 10.1016/j.csda.2004.06.015

Peter, B. M. (2022). A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis. *Philosophical Transactions of the Royal Society B*, *377*, 20200413. doi: 10.1098/rstb.2020.0413

R Core Team. (2022). *R: A language and environment for statistical computing (v4.1.3)*.

Rencher, A. C. (2002a). Discriminant Analysis: Description of Group Separation. In *Methods of Multivariate Analysis* (2nd ed., pp. 270–293). New York, USA: Wiley-Interscience.

Rencher, A. C. (2002b). Principal Component Analysis. In *Methods of Multivariate Analysis* (2nd ed., pp.

380–404). New York, USA: Wiley-Interscience.

Thia, J. A. (2022). Demonstration of the k − 1 limit for DAPC of genotype data. *Dryad*. doi: 10.5061/dryad.b8gtht7f0

Thia, J. A., & Riginos, C. (2019). *genomalicious*: serving up a smorgasbord of R functions for population genomic analyses. *BioRxiv*, 667337. doi: 10.1101/667337

Venables, W. N., & Ripley, B. D. (2002). *Statistics Complements to Modern Applied Statistics with S* (4th ed.). New York, USA: Springer.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., … Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wright, S. (1949). The genetical structure of populations. *Annals of Human Genetics*, *15*(1), 323–354.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36.

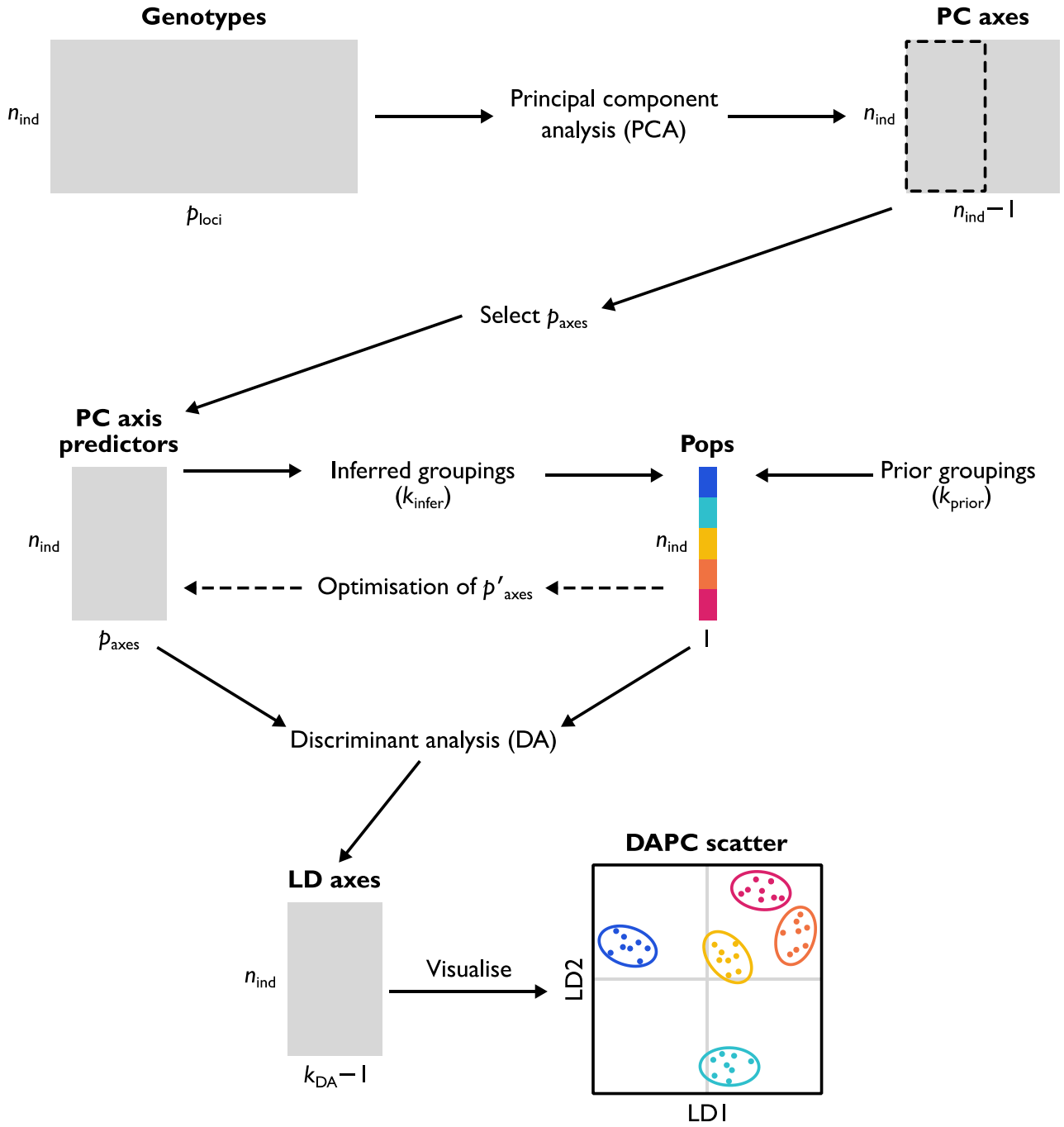


Figure 1. Schematic representation of a DAPC (discriminant analysis of principal components). A genotype matrix comprising n_{ind} samples genotyped at p_{loci} genetic loci is subjected to a PCA (principal component analysis). Assuming that $n_{ind} < p_{loci}$, there are $n_{ind} - 1$ PC axes with non-zero eigenvalues that summarise the allelic covariances in the sample. A subset of these PC axes, p_{axes}, are selected to represent population structure. Grouping of individuals into populations may be based on some *a pirori* expectation (k_{prior}), or may be inferred from the PCA (k_{infer}) using, for example, examination of the explained variance, PC scatter, or *K*-means clustering. Optionally, the *adegenet* functions XVALDAPC or OPTIM.A.SCORE can be used to produced an optimised number of predicted PC axes, p'_{axes}. A DA (discriminant analysis) is then used to identify linear combinations of the PC axes as predictors of differences among k_{DA} groups, where $k_{DA} = k_{prior}$ or $k_{DA} = k_{infer}$. Assuming $p_{axes} \geq k_{DA} - 1$, a set of $k_{DA} - 1$ LD (linear discriminant) axes is returned that summarise the among population variation. Visualisation of individuals projected into LD space is used to infer genetic relationships among individuals and populations.

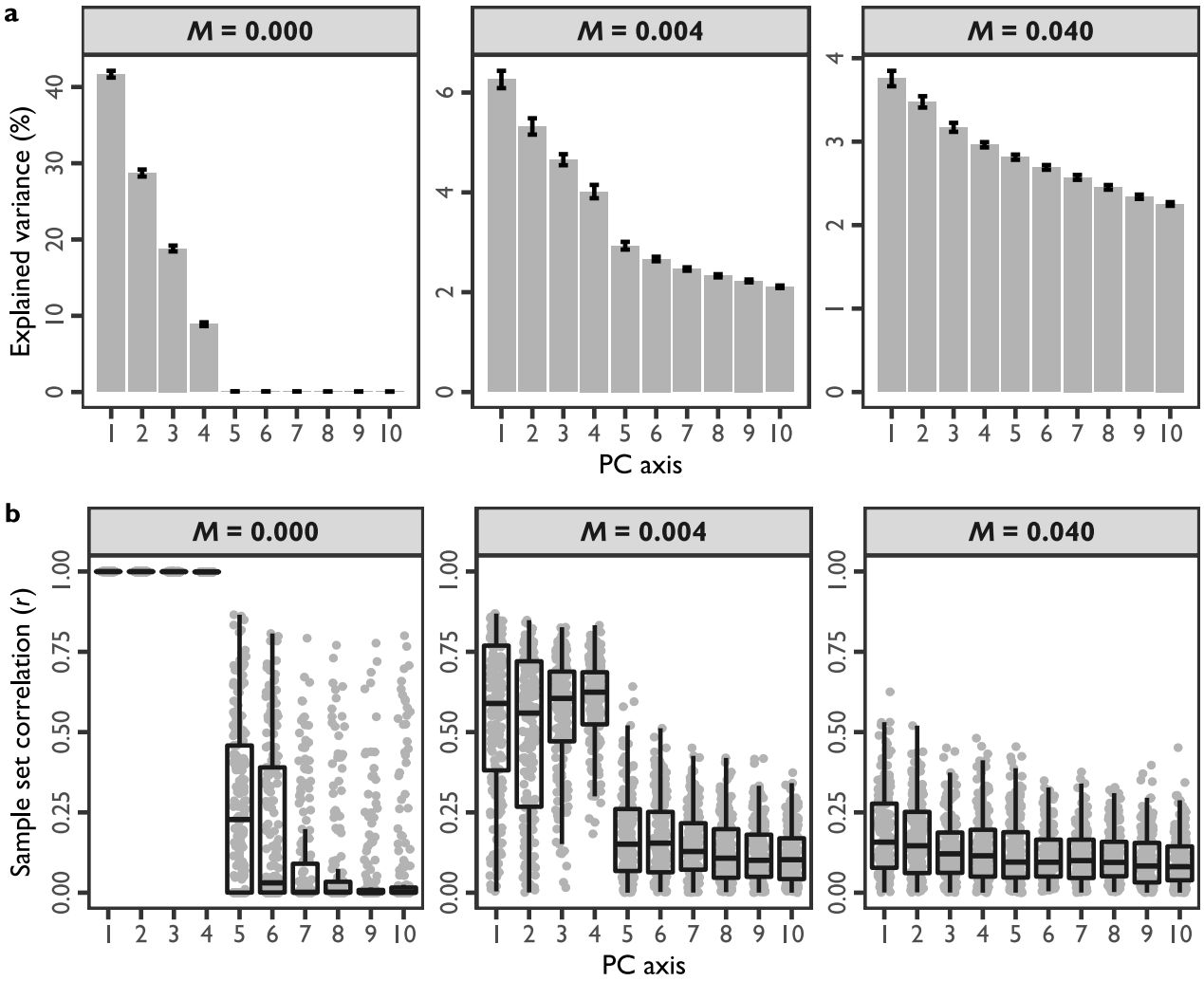


Figure 2. Principal component analysis under different migration scenarios. (a) Average explained variance for PC axes across simulated sample sets, using the *replicate simulation dataset*. The *x*-axis represents the first 12 PC axes. The *y*-axis is the amount of explained genotypic variance (percent of total) ascribed to each PC axis. Grey bars indicate the average across replicate simulations with the 95% confidence intervals. (b) Correlations among PC axis eigenvectors in the *singular simulation dataset*. The *x*-axis represents the first 10 PC axes. The *y*-axis represents the correlation of eigenvector loadings (contributions of each locus to variation on that PC axis). Each grey point represents a pair of independent sample sets from the same simulated metapopulation. Boxplots summarise the distribution of correlations across sample set pairs.

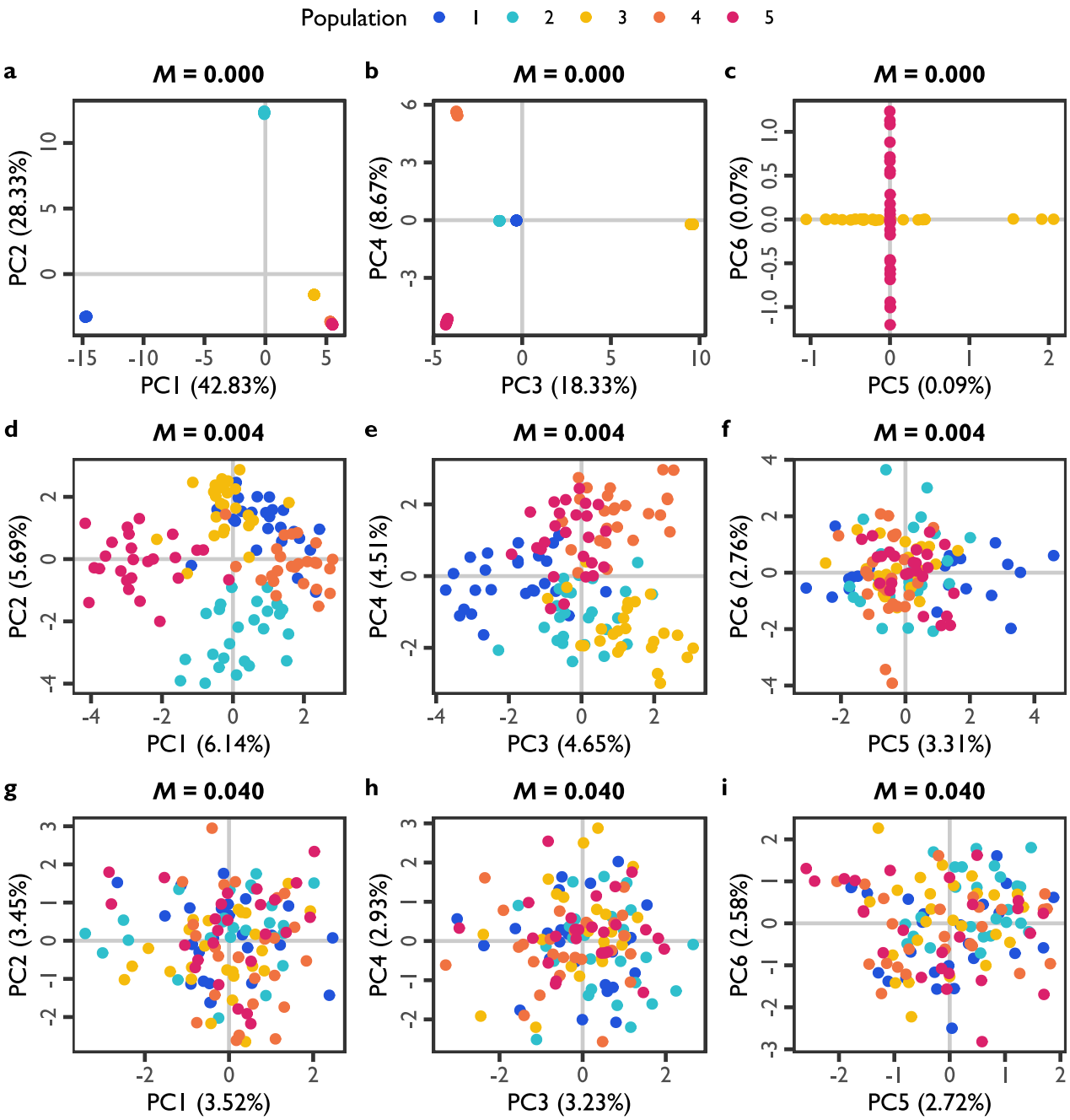


Figure 3. Exemplar scatterplots of individuals projected into PC space. The *x*-axis and *y*-axis, respectively, represent different pairs of PC axes. Numbers in parentheses indicate the amount of explained genotypic variance captured by each PC axis (percent of total). Each point represents an individual, coloured by population (see legend). Migration rates: (a, b, c) $M = 0.000$; (d, e, f) $M = 0.004$; and (g, h, i) $M = 0.040$. PC axis pairs: (a, d, g) 1 versus 2; (b, e, h) 3 versus 4; and (c, f, i) 5 versus 6.

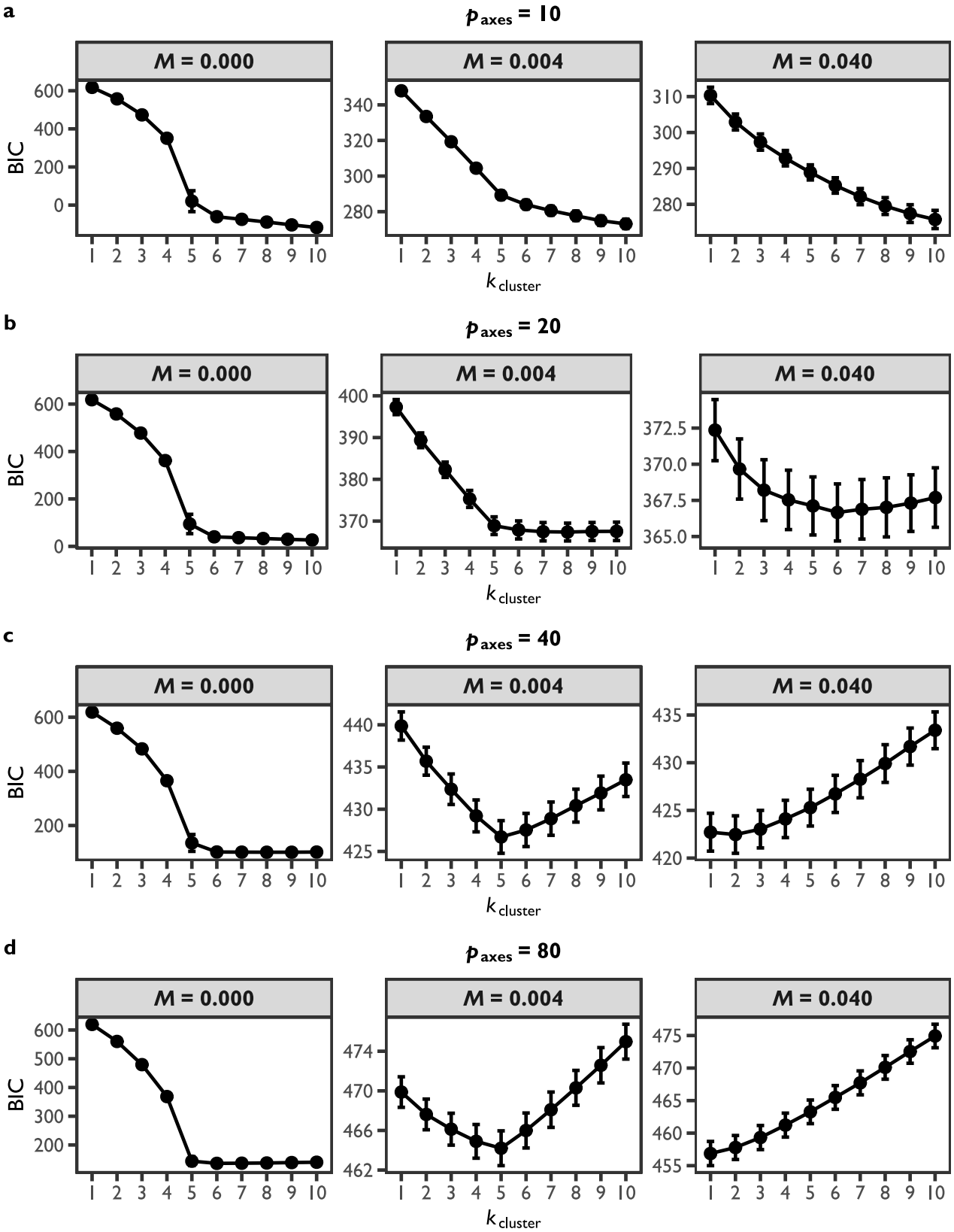
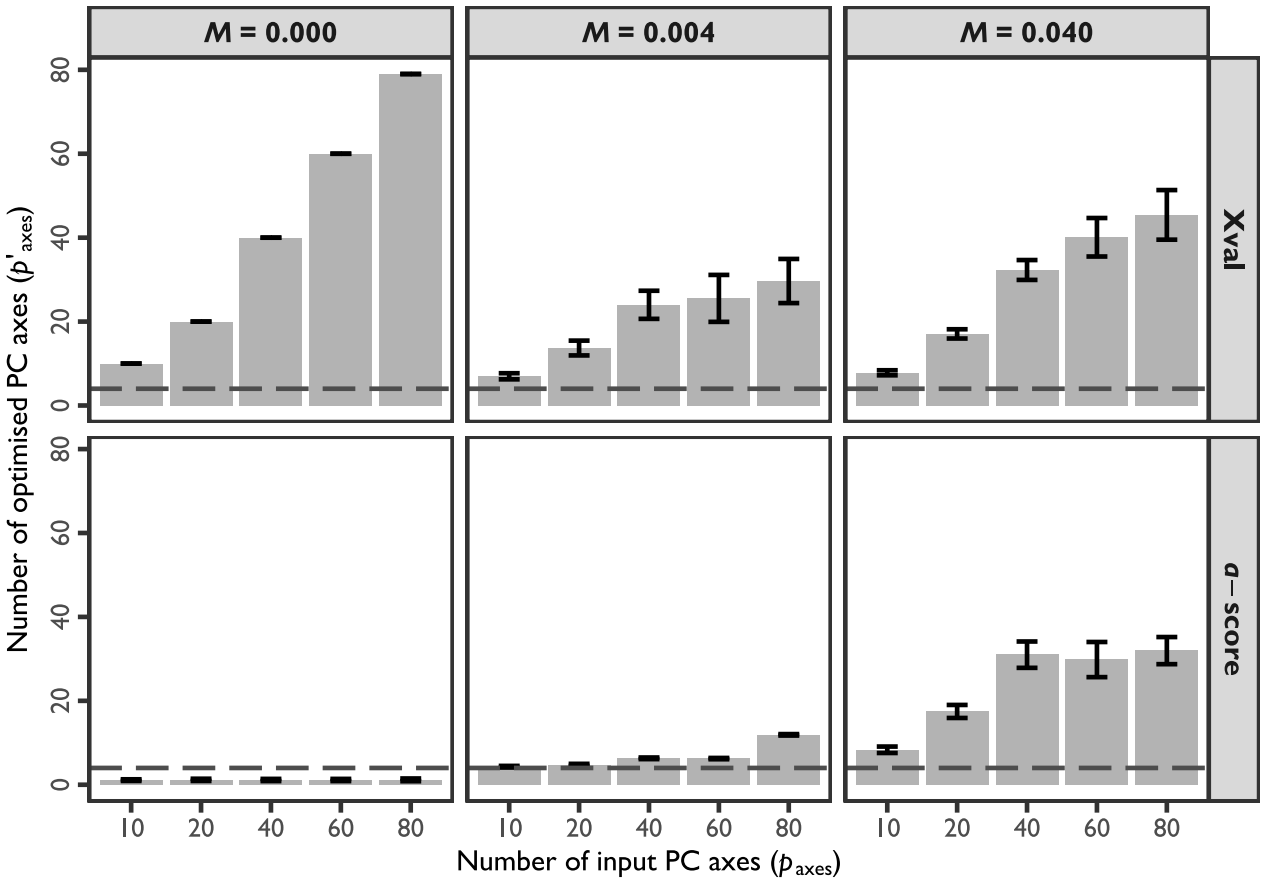


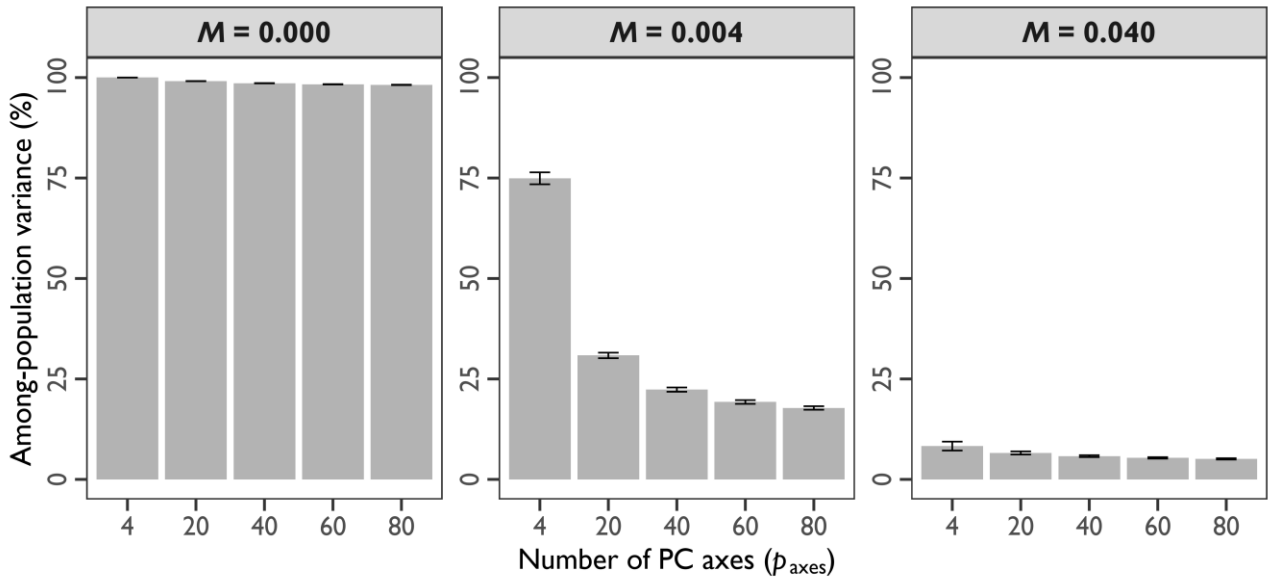
Figure 4. The inferred number of populations by *K*-means clustering, using the *replicate simulation dataset*. The *x*-axis is the number of clusters to divide samples among in *K*-means clustering, $k_{cluster}$. The *y*-axis is the associated BIC score (lower values are more likely). Points represent the mean BIC across replicate simulations with 95% confidence intervals. Panels represent each of the simulated migration scenarios. Number of initial PC axes, p_{axes}: (a) 10, (b) 20, (c) 40, and (d) 80.



503

504 **Figure 5.** Optimisation of PC axes to fit DA models using the *replicate simulation dataset*. The *x*-axis is the
505 number of PC axes used to initially fit the DA, p_{axes}. The *y*-axis is the number of optimised PC axes, p'_{axes},
506 using one of two optimisation algorithms available in the *adegenet* package. Panels represent combinations
507 of optimisation algorithm (in rows) and migration scenario (in columns). Bars represent the mean p'_{axes}
508 estimated across replicate simulations with 95% confidence intervals. The dashed lines demarcate $p'_{axes} = 4$.

509



510

511 **Figure 6.** Average amount of among-population variance captured on PC axes using the *replicate simulation*
512 *dataset*. The *x*-axis represents the total number of leading PC axes used in a DA, with k_{DA} = 5 hypothesised
513 groups. The *y*-axis represents the amount of variance (percent of total) that is among populations. Grey
514 bars indicate the average across replicate simulations with 95% confidence intervals. Panels contain results
515 for each migration scenario.

516

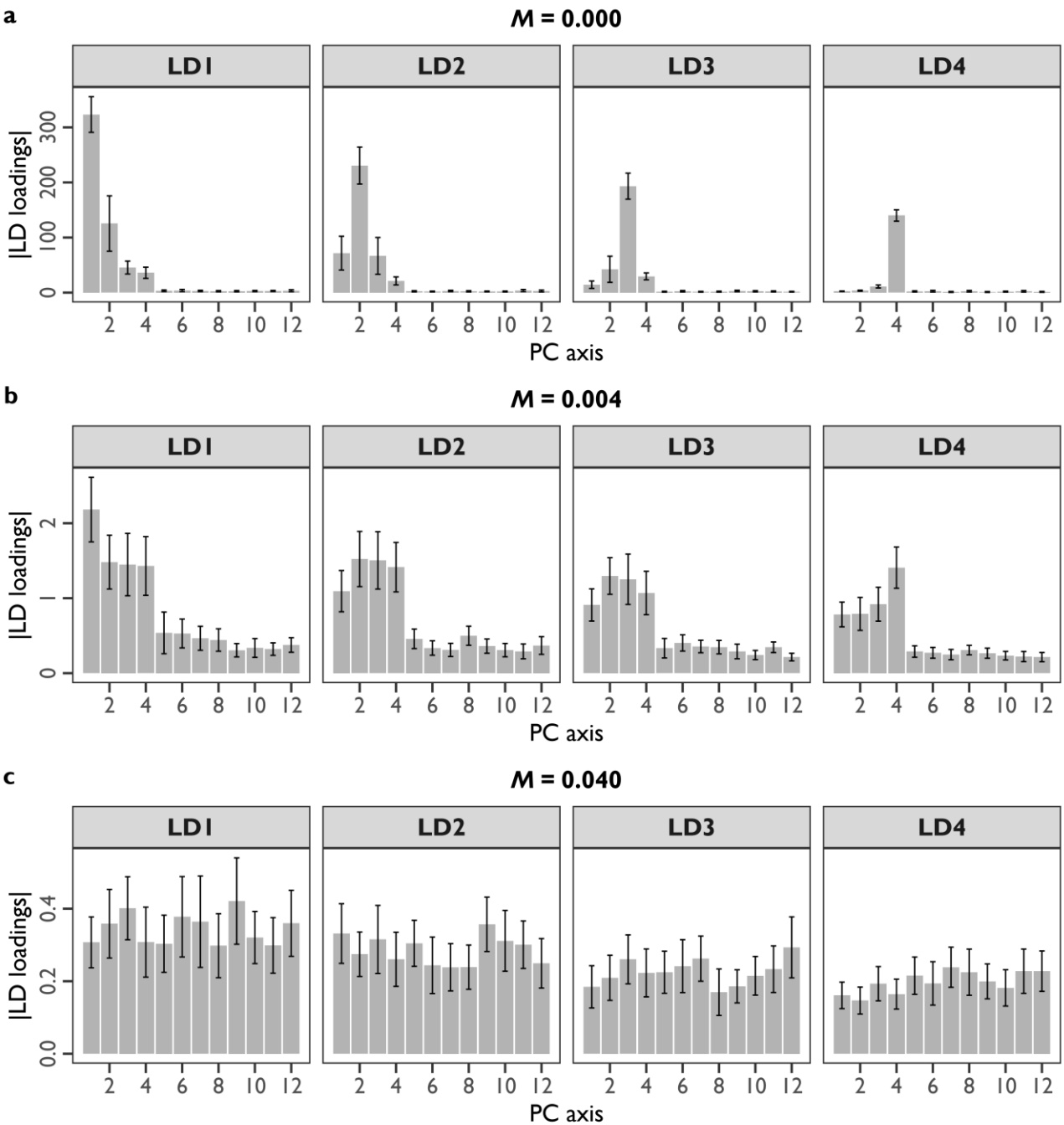


Figure 7. Average loadings of individual PC axes in DAs using the *replicate simulation dataset*. The *x*-axis represents the first 12 PC axes in DAs parameterised with k_{DA} = 5 and p_{axes} = 80 PC axes. The *y*-axis represents the absolute loadings (contributions to variation) of PC axes onto discriminant functions describing each of the four LD axes (panels). Grey bars indicate the average loadings across replicate simulations with 95% confidence intervals. Migration rates: (a) *M* = 0.000; (b) *M* = 0.004; and (c) *M* = 0.040.

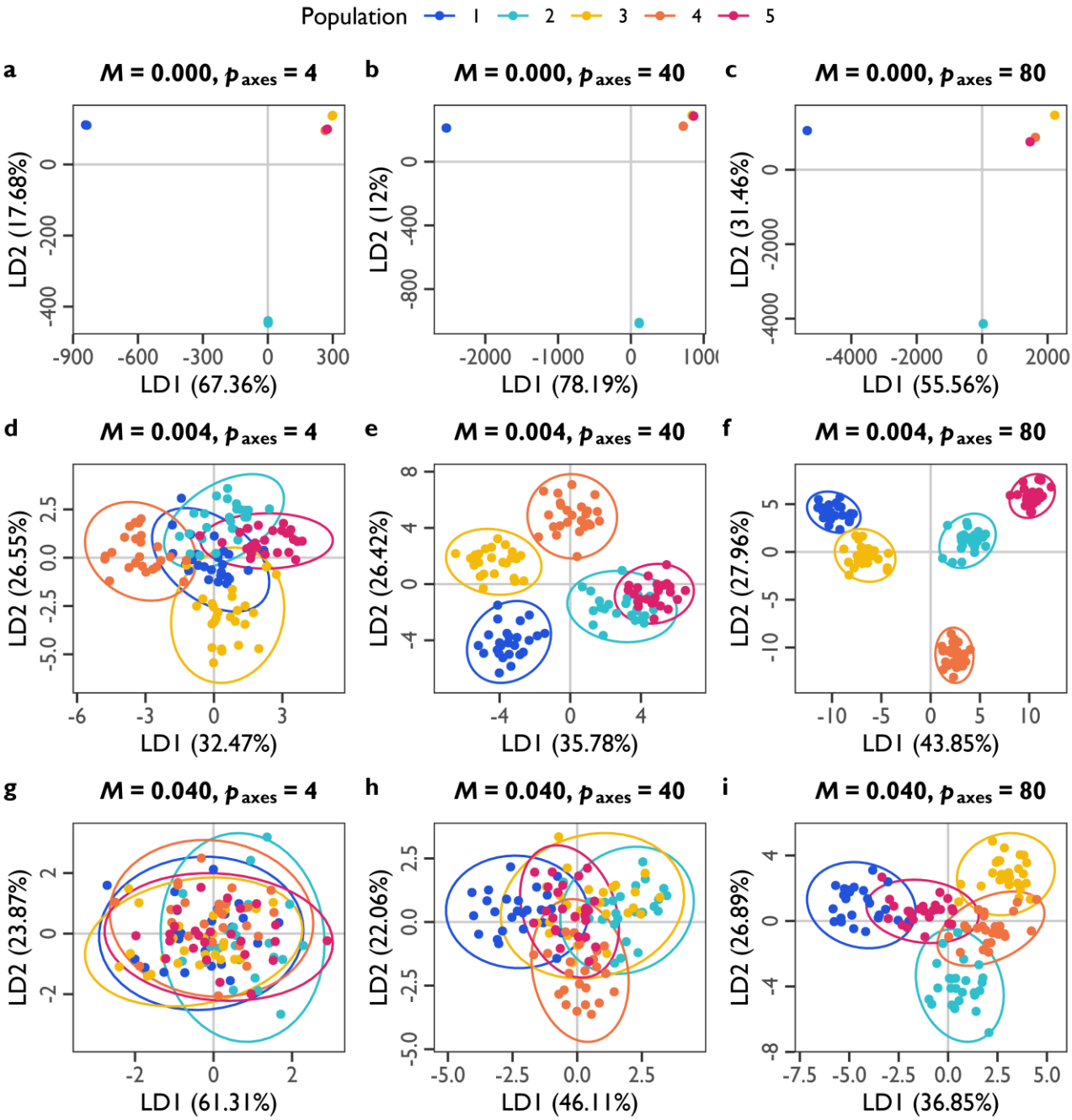


Figure 8. Exemplar scatterplots of individuals projected into LD space. The *x*-axis and *y*-axis represent the first and second LD axes, respectively, obtained from a DAPC parameterised with different numbers of PC axes, p_{axes}, with k_{DA} = 5 hypothesised groups. Numbers in parentheses indicate the amount of explained among-population variance for each LD axis (percent of total). Each point represents an individual, coloured by population (see legend). Migration rates: (a, b, c) M = 0.000; (d, e, f) M = 0.004; and (g, h, i) M = 0.040. Number of PC axes: (a, d, g) p_{axes} = 4; (b, e, h) p_{axes} = 40; and (c, f, i) p_{axes} = 80.

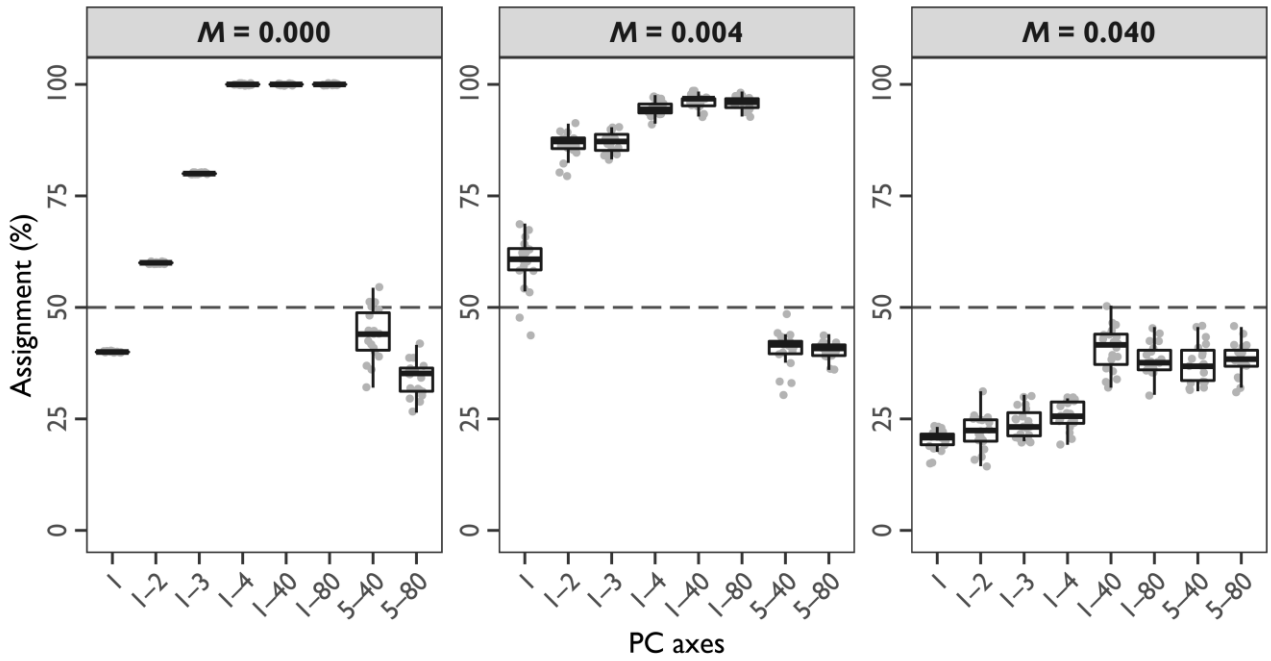


Figure 9. Correct assignment rates using DA models parameterised with different combinations of PC axes using the *singular simulation dataset*. DA models were constructed with a *training* sample set and used to assign populations to 19 *testing* sample sets, with a k_{DA} = 5. The *x*-axis represents the combinations of PC axes that either include the first four biologically informative *k* – 1 PC axes (1, 1–2, 1–3, 1–4, 1–40, and 1–80) or exclude them (5–40 and 5–80). The *y*-axis represents the percentage of correct assignment. Each grey point represents one of the *testing* sample sets. Boxplots summarise the distribution of correct assignment rates across *testing* sample sets. The dashed line demarcates a correct assignment rate of 50% (equally likely to correctly or incorrectly assign population). Migration rates: (a) *M* = 0.000; (b) *M* = 0.004; and (c) *M* = 0.040.

544 **Table 1.** Summary of mathematical notation.

Notation	Definition
k	The true number of effective populations in a genotype dataset.
k_{prior}	The *a priori* expectation of the number of populations, based on criteria external to the genotype data, for example, sampling locations.
k_{infer}	The inferred number of effective populations, an estimate of k, from genotype data.
$k_{cluster}$	The number of clusters to divide sample among in K-means clustering. Many values could be tested. The most likely value of $k_{cluster}$ can be used as inference of the number of effective populations, k_{infer}.
k_{DA}	The number of hypothesised groups to discriminate among in a DAPC, a researcher's choice of k_{prior} or k_{infer} to analyse.
p_{axes}	The number of leading PC axes to use as predictors of differences among k_{DA} groups in a DAPC.
p'_{axes}	The optimised number of leading PC axes to use as predictors of differences among k_{DA} in a DAPC.
m	The individual migration rate from a source population into a focal population.
M	The total migration rate into a focal population, the summed m across all source populations.

545

546

Supplementary: Analytical methods

F_{ST} calculations

Estimates of Weir and Cockerham's F_{ST} were performed using the FSTWC_GENOS function from the R package, *genomalicious*.

Eigenvector correlations

Eigenvector correlations were determined between pairs of independent sample sets in the *singular simulation dataset*. For each sample set, I performed a PCA on the allelic covariances using R's PRCOMP function, with the arguments "scale=FALSE, center=TRUE". For each ith eigenvector from 1–10, I calculated the Pearson correlation between each jth sample set pair, using R's COR function. This allowed me to determine how similar each ith eigenvector was among the different sample sets.

Discriminant analysis of principal components

Multiple DAPCs were performed on independent simulations from the *replicate simulation dataset.* The *genomalicious* function ADEGENET_DT2GENLIGHT was used to convert genotype data tables into *adegenet* 'genlight' objects. The DAPC function in *adegenet* was then used to fit DAPCs using p_{axes} = 4, 20, 40, 60 and 80 PC axes.

Among-population variation

The proportion of among-population variation captured by each value of p_{axes} in the *replicate simulation datasets* was calculated by performing a multivariate analysis of variance (MANOVA). I used R's MANOVA function to fit the model:

$$\mathbf{Y} = \text{POPULATION}$$

Where **Y** is a matrix of PC axes amd POPULATION is a categorical factor of population identity. I used the SUMMARY function on the 'manova' object to extract the sum of squares and cross products matrices for the hypothesis and error term, **H** and **E**, respectively. The proportion of among-population variation was then calculated as:

$$trace\{\mathbf{H}\} / (trace\{\mathbf{H}\} + trace\{\mathbf{E}\})$$

Where *trace*{} indicates the trace of the associated matrices.

K-means and optimisation of input PC axes

For p_{axes} = 10, 20, 40, 60 and 80 PC axes, I estimated k_{infer} and tested the performance of optimisation algorithms to derive an optimised number of PC axes, p'_{axes}, that matched the $k - 1$ expectation. These analyses were conducted on simulations from the *replicate simulation dataset*.

K-means clustering was performed using R's KMEANS. For each value of p_{axes}, I tested support for k = 1–10 clusters. A BIC score was calculated as support for each value of k.

Optimisation to derive p'_{axes} was conducted using the XVALDAPC and OPTIM.A.SCORE functions from the *adegenet* package. For each function and input p_{axes} combination, I tested all possible 1–p_{axes} number of PC axes, each with 50 iterations.

582 ## Population assignment

583 Population assignment analyses were conducted using the *singular simulation dataset*. One of the
584 independent sample sets was used as a *training* sample set, with the remaining 19 sample sets used as
585 *testing* sample sets. PCA of allelic covariances in the *training* sample set was conducted using the PRCOMP
586 function with the arguments "scale=FALSE, center=TRUE". A series of DAPC models were manually fit using
587 different combinations of PC axes from the *training* sample sets: 1, 1−2, 1−3, 1–4, 1–40, 1–80, 5–40, and 5–
588 80 PC axes. First, the LDA function from the *MASS* package was used to perform a DA using these different
589 combinations of PC axes. Second, individuals from each *testing* dataset were projected into the same PC
590 space as the *training* sample set, using eigenvectors from the *training* sample set PCA. Third, the DA model
591 constructed on the *training* sample set was used to project individuals from the *testing* sample sets into the
592 *training* sample set's LD space. Finally, population assignment was performed using R's PREDICT function.

593 ## Supplemental simulations of more complex demographic scenarios

594 For the purpose of illustrating the k – 1 consideration under more complex demographic scenarios, I
595 modelled isolation-by-distance and hierarchical structure using *fastsimcoal2*. Similar to the core
596 simulations, I modelled five populations following a series of sequential divergence events (Figure S1), with
597 each population containing 500 diploid individuals, and genome with a length of 20 Mb, a recombination
598 rate of 1e−8, and a mutation rate of 2e−9. At the end of the simulation, I sampled 25 diploids from each
599 population.

600 For the isolation-by-distance simulation, I modelled migration between adjacent populations in a sequential
601 array: 1, 2, 3, 4, 5. Adjacent populations were connected by a pairwise migration rate of m = 0.005.
602 Therefore, populations with two neighbouring populations (2, 3 and 4) had a total migration rate of M=
603 0.010, and those at the ends of the population array (1 and 5) had a total migration rate of M = 0.005.

604 For the hierarchical structure simulation, I modelled two levels of migration: within and between groups.
605 Populations were split into two groups, with populations 1 and 2 forming group 1, and populations 3, 4 and
606 5 forming group 2. Pairwise migration rates were m = 0.0025 between populations within regions, whereas
607 populations between regions were connected by a pairwise migration rate of m = 0.0005.

608 SNP genotypes were imported into R. A PCA of allelic covariances was performed for each demographic
609 scenarios using all SNPs. *K*-means clustering was used to assess the number of effective populations from k
610 = 1 to k = 10. BIC scores were calculated as the test statistic for each value of k.

611

Supplementary: Figures

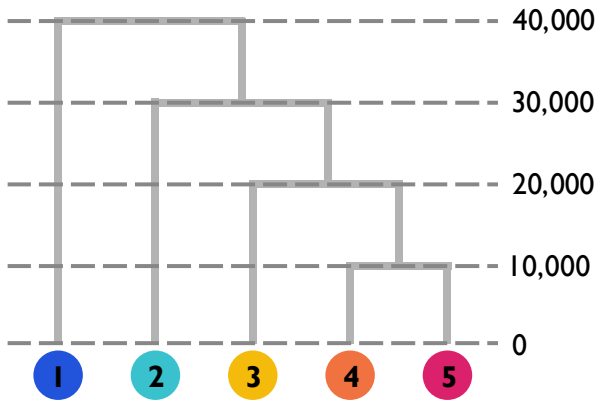


Figure S1. Simulated population divergence times. Each population (1–5) was comprised of 500 diploid individuals. Annotated nodes demarcate the number of generations in the past for each divergence event.

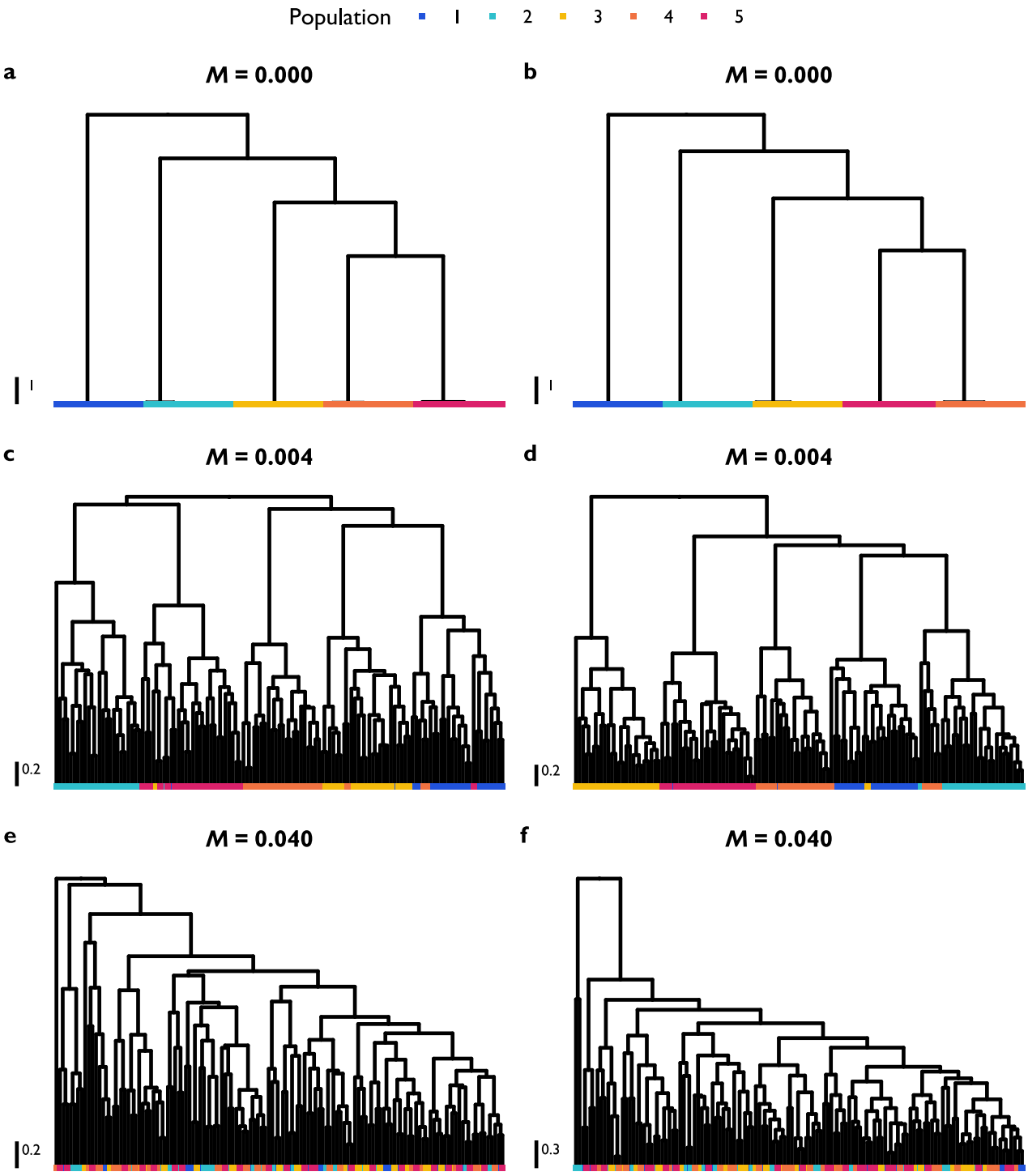


Figure S2. Hierarchical clustering of simulated samples in PC space from the *replicate simulation dataset*. Euclidean distance was estimated between all pairs of individuals using PC axes 1–4. Distance matrices were subjected to hierarchical clustering using the UPGMA method (unweighted pair group method with arithmetic mean) in R's HCLUST function from the *stats* package. Branch lengths depict the Euclidean distance among individuals (see scalebar). Tips represent individuals, coloured by population (see legend). Examples of replicate simulations for each migration scenario: (a–c) M = 0.000; (d–f) M = 0.004; and (g–i) M = 0.040.

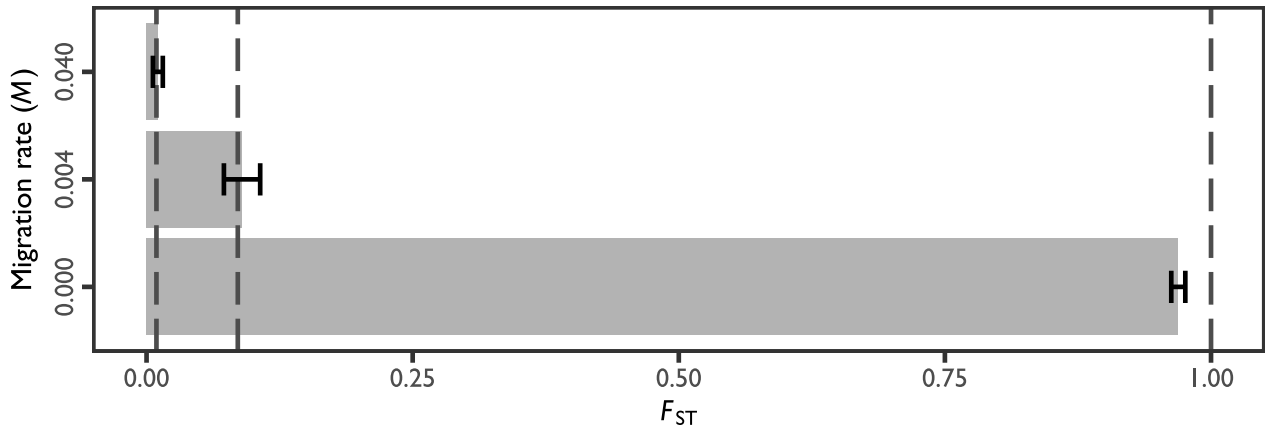
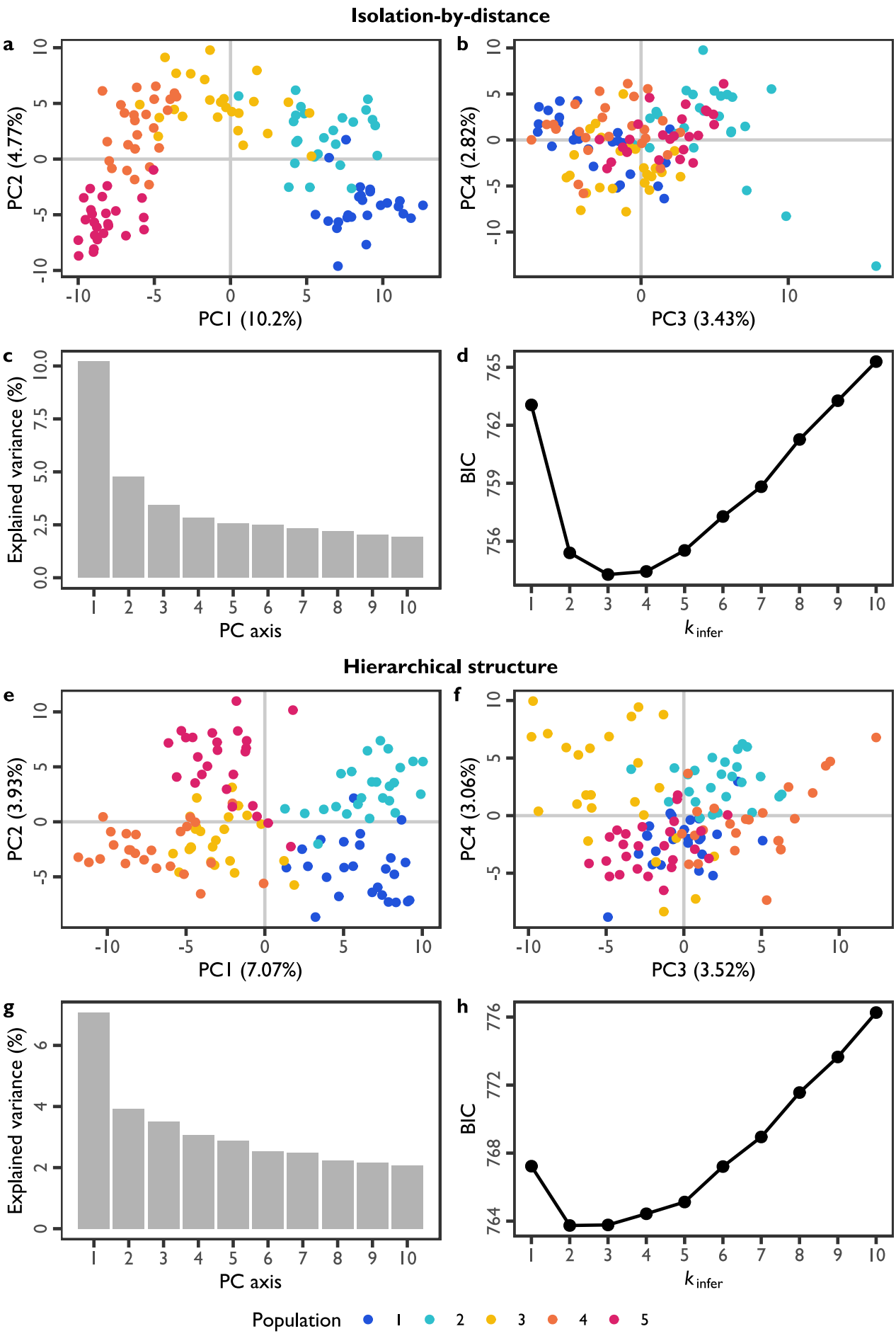


Figure S3. F_{ST} among populations in the *replicate simulation dataset*. The *y*-axis is the simulated migration rates and the *x*-axis is the estimated F_{ST}. Grey bars depict the mean F_{ST} across replicate simulations with 95% confidence intervals. Dashed lines demarcate the expected F_{ST} values at each migration rate (from left to right): F_{ST} = 0.01, 0.09, and 0.99, for M = 0.040, 0.004, and 0.000, respectively (Takahata & Nei, 1984).


Isolation-by-distance
a
PC2 (4.77%)
PC1 (10.2%)
b
PC4 (2.82%)
PC3 (3.43%)
c
Explained variance (%)
PC axis
d
BIC
k_{infer}
Hierarchical structure
e
PC2 (3.93%)
PC1 (7.07%)
f
PC4 (3.06%)
PC3 (3.52%)
g
Explained variance (%)
PC axis
h
BIC
k_{infer}
Population 1 2 3 4 5

634

Figure S4. Population structure under more complex demographic scenarios. (a–d) Isolation-by-distance. (e–h) Hierarchical structure. (a, b, e, f) PC scatter plots. The *x*-axis and *y*-axis, respectively, represent different pairs of PC axes. Numbers in parentheses indicate the amount of explained genotypic variance captured by each PC axis (percent of total). Each point represents an individual, coloured by population (see legend). (c, g) Scree plots of explained variance. The *x*-axis represents the first 10 PC axes. The *y*-axis is the amount of explained genotypic variance (percent of total) ascribed to each PC axis. (d, h) Predicted number of populations using *K*-means. The *x*-axis is the inferred number of clusters using p_{axes} = 80 PC axes. The *y*-axis is the associated BIC score (lower values are more likely).