DAPC of genotype data only requires the first $k - 1$ PC axes for k populations

Joshua A. Thia [1, *]

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

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

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 value of p_{axes} is a deterministic feature of a genotype dataset based on population genetic theory. For k discrete populations, only the first $k - 1$ PC axes are expected to be biologically informative and capture population structure. DAs fit using more than the first $k - 1$ PC axes are over-parametrised and may discriminate groups using biologically uninformative predictors. Using samples drawn from simulated metapopulations, I show that DAPCs parameterised with the appropriate $k - 1$ 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.

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 are 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 first 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* discrete 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). DA takes a set of measured *p* predictor variables and identifies linear combinations of those variables that maximise the discrimination of individuals from predefined 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, and then uses the resulting PC axes as predictors in a DA model (Jombart et al., 2010) (Figure 1).

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 analyses can be made more transparent by specific reporting of how groups are predefined, clearly defining the goals of a DAPC, and stating the number of PC axes used in the analysis.

In this piece, I specifically address the consideration of the number of PC axes used to construct the DA model in a DAPC of population structure relative to the number of *k* populations. Critical to a DA are decisions on: (1) how many *k* groups to discriminate among; (2) which individuals are allocated to each group; and (3) what variables will be used to discriminate the groups. The value of *k* can be defined based on some *a prior* expectation, k_{prior}, such as the number of sampled locations or biogeography. Alternatively, *k* can be inferred directly from the genetic data, k_{infer}, using various approaches. Either way, when discriminating among *k* groups in a DA model, we are making an explicit hypothesis about how variation is organised among *k* groups based on a set of predictor variables and our allocation of individuals to those *k* groups.

DAPCs of population structure are often parameterised with many PC axes, p_{axes}, with the intention of including a substantial proportion of the total genotypic variance in the DA model. Yet based on the expectation of $k - 1$ PC axes for k populations (Patterson et al., 2006), the use of $p_{axes} > k - 1$ is over-parameterised. Here, I propose that a $k - 1$ limit sets a deterministic limit on the maximal number of PC axes that should be used in a DAPC. 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 first $k - 1$ PC axes, and that inclusion of many more than $k - 1$ PC axes inflates the perceived differences among groups. These results indicate that appropriate parameterisation of a DAPC would benefit from attempts to match the input p_{axes} with the likely k in genotype datasets.

Simulated genetic datasets

I simulated metapopulations comprising five discrete 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 recombination between adjacent bases at 1e−8 and a mutation rate of 2e−8. Populations underwent a series of hierarchical divergence events, each 10,000 generations in the past (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).

The modelled scenarios reflect different contexts where the number of populations may or may not be equal to the effective k. 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 weak, but there is still a discernible k = 5. In the M = 0.040 scenario, there is no population structure, and the effective k in this case is k = 1. To illustrate how mismatching k with the choice of p_{axes} influences a DPAC, I will intentionally fit DA models with k = 5 in all scenarios. This choice is appropriate for M = 0.000 and M = 0.004, but inappropriate for M = 0.040. The purpose of this exercise is to demonstrate how the hypothesis driven nature of a DA and inappropriate parameterisation can produce results that are misleading with respect to the true population structure.

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 different simulated metapopulations that had undergone the same migration scenario.

Simulations 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 variable 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 covariances among a set of variables presents a major challenge for the use of these PC axes in predictive analyses (Peres-Neto, Jackson, & Somers, 2005). For genotype data, Patterson, Price, & Reich (2006) demonstrated a clear expectation that for k discrete populations there are $k - 1$ PC axes that capture among-population variation, with all remaining axes characterising variation within populations. Whilst there are different methods to test the number of biologically informative PC axes (Cattell, 1966; Jackson, 1993; Peres-Neto et al., 2005; Shriner, 2011), these methods and their performance under different population genetic scenarios or study designs are not the focus of this present work. Instead, what I wish to highlight is that only a limited number of PC axes ($k - 1$) capture biologically meaningful variation, and it is only these axes that should be used to parameterise a DAPC.

Aside from statistical tests, a simple way to assess the number of biologically informative PC axes is to examine the scree of eigenvalues and their explained variance for each PC axis. When population structure exists, there is typically an inflection point in the scree plot that creates an "elbow shaped" distribution of explained variances (Abegaz et al., 2019; Cattell, 1966). This inflection point is usually indicative of k, 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, relative to the $M = 0.004$ scenario. Contrastingly, in the $M = 0.040$ scenario, although there are five populations, there is no variation among them, and the scree slope exhibits a smooth decline. In the $M = 0.040$ scenario, effectively, $k = 1$. Scree plots can therefore be used to infer the putative genetic groups in a sample, k_{infer} (Figure 1).

Expectations of the number of effective k populations from screen plots (Figure 2a) are reflected in the estimated number of genetic clusters inferred using K-means clustering (Figure 2b) and from examination of scatter plots of individuals projected into PC space (Figure 3). As we can see, the K-means inferred genetic clusters match expectations of the effective k from scree plots (compare Figure 2a with 2b; see also Figure S4). In the $M = 0.000$ scenario, BIC scores rapidly drop and plateau at $k_{infer} = 5$, and for $M = 0.004$, there is a clear inflection at $k_{infer} = 5$. However, for $M = 0.040$, BIC is smallest at $k_{infer} = 1$. In PC space, we can see that where population structure exists it is only found on the first $k - 1$ PC axes. 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. At this point ($\geq k$), we are no longer visualising genotypic variation among populations, but instead, variation that is within populations. In contrast, in $M = 0.040$, there is no clustering of individuals in PC space on the first four PC axes because there is no population structure and the effective $k = 1$. The key point here is that based on the number of distinct genetic groups, k, there are $k - 1$ PC axes that are biologically informative and those PC axes $\geq k$ are biologically uninformative.

Another way to illustrate the biological relevance of the first $k - 1$ PC axes is with respect to the recoverability of eigenvectors among independent sample sets from the same populations. 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. Eigenvectors of the first $k - 1$ PC axes will be correlated across independent sample sets, whereas eigenvectors of $\geq k$ PC axes will not be 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. Again, when there is population structure, the first $k - 1$ PC axes capture the informative variation that exists among populations. Contrastingly, PC axes $\geq k$ capture stochastic sampling noise in structured populations.

In summary, the analyses presented in this section demonstrate that only the first $k - 1$ PC axes for k populations are biologically informative. Knowledge that only the first $k - 1$ PC axes capture population structure can be used to validate prior assumptions of genetic groupings (k_{prior}), or to help infer the probable k when k is unknown (k_{infer}). Whereas PC axes $\leq k - 1$ capture among-population variation and describe repeatable linear combinations of loci across independent sample sets, PC axes $\geq k$ capture the sampling variance within populations and are not replicable across independent sample sets.

Considerations for discriminant analysis of principal components

That we only expect the first $k - 1$ PC axes to capture population structure among k discrete genetic groups implies that the number of PC axes used in the DA portion of a DAPC should be $p_{axes} = k - 1$. Put another way, because DA is specifically aiming to maximise among-population variation, the variables provided should ideally be biologically informative, such that the solution produced is biologically meaningful. However, parameterisation of DAPC solely using the first $k - 1$ PC axes is not commonplace in population genetic studies. Instead, typical parameterisation often involves setting p_{axes} to capture a certain proportion of the variance in genotypes. Alternatively, the XVALDAPC and OPTIM.A.SCORE functions in the *adegenet* package (Jombart, 2008; Jombart & Ahmed, 2011) provide built-in automated methods to estimate the number of optimised PC axes to use, 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. Based on analyses using the *replicate simulation dataset*, the XVALDAPC

function consistently returned a p'_{axes} that was greater than the $k - 1 = 4$ expectation across all migration scenarios and the initial p_{axes}. The OPTIM.A.SCORE function tended to be more conservative, but increasing migration rates and a larger p_{axes} led to inflated estimates of p'_{axes}.

Because population structure is only captured on the first $k - 1$ PC axes, the addition of axes beyond $k - 1$ does not increase the variation among populations. Figure 5 depicts the mean variation among populations when using different values of p_{axes} to parameterise a DAPC with the hypothesis that $k = 5$. 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 $M = 0.001$ 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}. Using $k - 1$ PC axes to parameterise the DA portion of a DAPC is more parsimonious, uses only biologically relevant predictors, and maximises the total variation among populations in the input predictors.

Our expectation is that PC axes capturing population structure will contribute most to discrimination among populations. But as p_{axes} increases beyond $k - 1$, the DA solution can potentially be influenced by biologically uninformative predictors. Consider the contributions of PC axes in a highly over-parameterised DA models using $p_{axes} = 80$ PC axes. Figure 6 depicts the average absolute loadings of the first 12 PC axes onto the discriminant functions that describe the four-dimensional LD space when the hypothesis is $k = 5$. For the $M = 0.000$ and $M = 0.004$ where the effective k is indeed $k = 5$, PC axes 1–4 ($\leq k - 1$) 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 the DAs 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 $M = 0.000$, the DAs do not simply return the 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, there is no clear difference in the contributions of PC axes $\leq k - 1$ versus PC axes ≥ 5. For $M = 0.040$, all the variation is within populations, but the DAs derive solutions using all the predictor PC axes regardless, despite none of these 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 initial choice of 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 $k = 5$; 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 recover what is essentially 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 an effective $k = 1$ 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 very little 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 model fit. Hence, the degree to which a mismatch between the specified k and the choice of p_{axes} affects our interpretation of population structure from a DA may depend on: (1) the magnitude of genetic differentiation; and (2) the divergence of p_{axes} from the effective k in a genotype dataset.

Because DAs are 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 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 a $k = 5$ hypothesis, 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 4 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 first $k – 1$ PC axes. Exclusion of the first four $k – 1$ 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 (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 $\geq k$ are stochastic across different sample sets and are uninformative predictors for population assignment. The $M = 0.000$ scenario provides an extreme case where the effective $k = 1$, 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 a more appropriate parameterisation of DAPC is one where $p_{axes} = k – 1$. The addition of many more PC axes greater than $k – 1$ does not add informative among-population variation into the set of predictors used in a DA. Moreover, when genetic differentiation is weak, the inclusion of many uninformative PC axes produces a model that is over-fit and can give a false sense of the magnitude of among-population differences. DAs fit with the appropriate $k – 1$ PC axes for k populations are maximised for the among-population variation in the predictor variables, are not influenced by the idiosyncrasies of uninformative predictors, and are generalisable to independent sample sets.

Concluding remarks

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

et al., 2020). Because PC axes ≤ k – 1 are the only dimensions capturing population structure for k populations (Patterson et al., 2006), I suggest they are the only PC axes that should be used in a DAPC. In other words, the number of PC axes used in DAPC parameterisation is a deterministic value of a genotype dataset. Other population genetic methods that implement PCAs perform a selection step to limit analyses to those axes that capture population structure (for example: Conomos, Reiner, Weir, & Thornton, 2016; Luu, Bazin, & Blum, 2017; Meisner & Albrechtsen, 2018). Nonetheless, such considerations have never been discussed in the context of performing a DAPC, at least to the best of my knowledge.

DAPCs are frequently used for the visualisation of population structure. Hence, the projection of individuals into LD space is an important focal point for interpretations of the genetic relationships among populations. Inclusion of many more than k – 1 PC axes over-parameterises DA models with variables that do not capture biologically meaningful variation. The risk of this over-parameterisation is that it can give the impression of greater population structure through the appearance of tighter, more discrete clustering of individuals in LD space. Over-parameterisation of DAs is potentially most problematic in cases where genetic differentiation is weak (Figures 7c versus 7f versus 7i). When genetic differentiation is weak, the relative difference in contributions of biologically informative versus uninformative PC axes to the DA solution is smaller (Figures 6a versus 6b versus 6c). Using $p_{axes} = k - 1$ PC axes in a DAPC not only returns a solution that is closer to the true biology but fits a more parsimonious model that is less subject to the idiosyncrasies of uninformative predictors. Moreover, automated selection of PC axes to use in a DAPC, for example, with *adegenet*'s XVALDAPC and OPTIM.A.SCORE functions, may lead to considerable over-parametrisation under certain conditions (Figure 4). A k – 1 limit for p_{axes} therefore provides a straightforward, biologically based criteria for DAPC paramterisation.

In this work, I used a very simple simulated metapopulation for demonstrating the consideration of k – 1 PC axes for k discrete populations in a DAPC of genotype data. However, in empirical population genetic datasets, multiple factors might contribute to structuring of genotypes, such as demography, sampling effort, the number of genetic markers, and linkage disequilibrium among genetic markers (Abegaz et al., 2019; McVean, 2009; Shriner, 2011). Patterson et al. (2006) provided a test based on expectations under a Tracy-Widom distribution, but various statistical approaches, so called "stopping rules", exist to assess the number of informative PC axes in a dataset (Cattell, 1966; Jackson, 1993; Peres-Neto et al., 2005; Shriner, 2011). Prior to performing a DAPC, the putative informative PC axes should be scrutinised to ensure they capture structure among populations and not groups of tightly linked loci, such as sex chromosomes or inversions (Huang, Andrew, Owens, Ostevik, & Rieseberg, 2020; Shakya, Wang-Claypool, Cicero, Bowie, & Mason, 2022; Tepolt, Grosholz, de Rivera, & Ruiz, 2021). Loci in linkage groups should be removed, subset, or recoded into a single "super locus", to ensure they do not inflate the number of informative axes relative to k – 1 for k populations.

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 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, use of $p_{axes} = k - 1$ 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.

To conclude, I provide some suggestions as to how the considerations discussed in this work can be applied to analyses of genotype data in population genetic studies.

1. Where there is no *a priori k* and k_{infer} is used in a DAPC, limit discrimination to $p_{axes} = k_{infer} - 1$ PC axes.
2. Where there is an *a priori k* and k_{prior} is used in a DAPC, it is beneficial to determine k_{infer} because the effective *k* might differ from k_{prior}. When there is a mismatch, it may still be within a researcher's interest to test how variation is partitioned among k_{prior} groups, but the choice of p_{axes} should be made with respect to the relationship between k_{prior} and k_{infer}.
3. When $k_{prior} = k_{infer}$, limit discrimination to $p_{axes} = k_{prior} - 1$ PC axes.
4. When $k_{prior} < k_{infer}$, 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.
5. When $k_{prior} > k_{infer}$, limit discrimination to $p_{axes} = k_{infer} - 1$ PC axes. This choice ensures that only biologically informative predictors are used.
6. Check that k_{infer} makes sense. Various factors unrelated to population subdivision might contribute to structure in a genotype dataset. Understanding these factors is necessary for correctly inferring the effective number of *k* populations and selecting the appropriate p_{axes} for a DAPC.

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.

Huang, K., Andrew, R. L., Owens, G. L., Ostevik, K. L., & Rieseberg, L. H. (2020). Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *Molecular Ecology*, *29*(14), 2535–2549. doi: 10.1111/mec.15428

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

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.

Shakya, S. B., Wang-Claypool, C. Y., Cicero, C., Bowie, R. C. K., & Mason, N. A. (2022). Neo-sex chromosome evolution and phenotypic differentiation across an elevational gradient in horned larks (Eremophila Alpestris). *Molecular Ecology*.

Shriner, D. (2011). Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity*, *107*(5), 413–420. doi: 10.1038/hdy.2011.26

Tepolt, C. K., Grosholz, E., de Rivera, C., & Ruiz, G. (2021). Balanced polymorphism fuels rapid selection in an invasive crab despite high gene flow and low genetic diversity. *Molecular Ecology*, 1–15. doi: 10.1111/mec.16143

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.

Figures

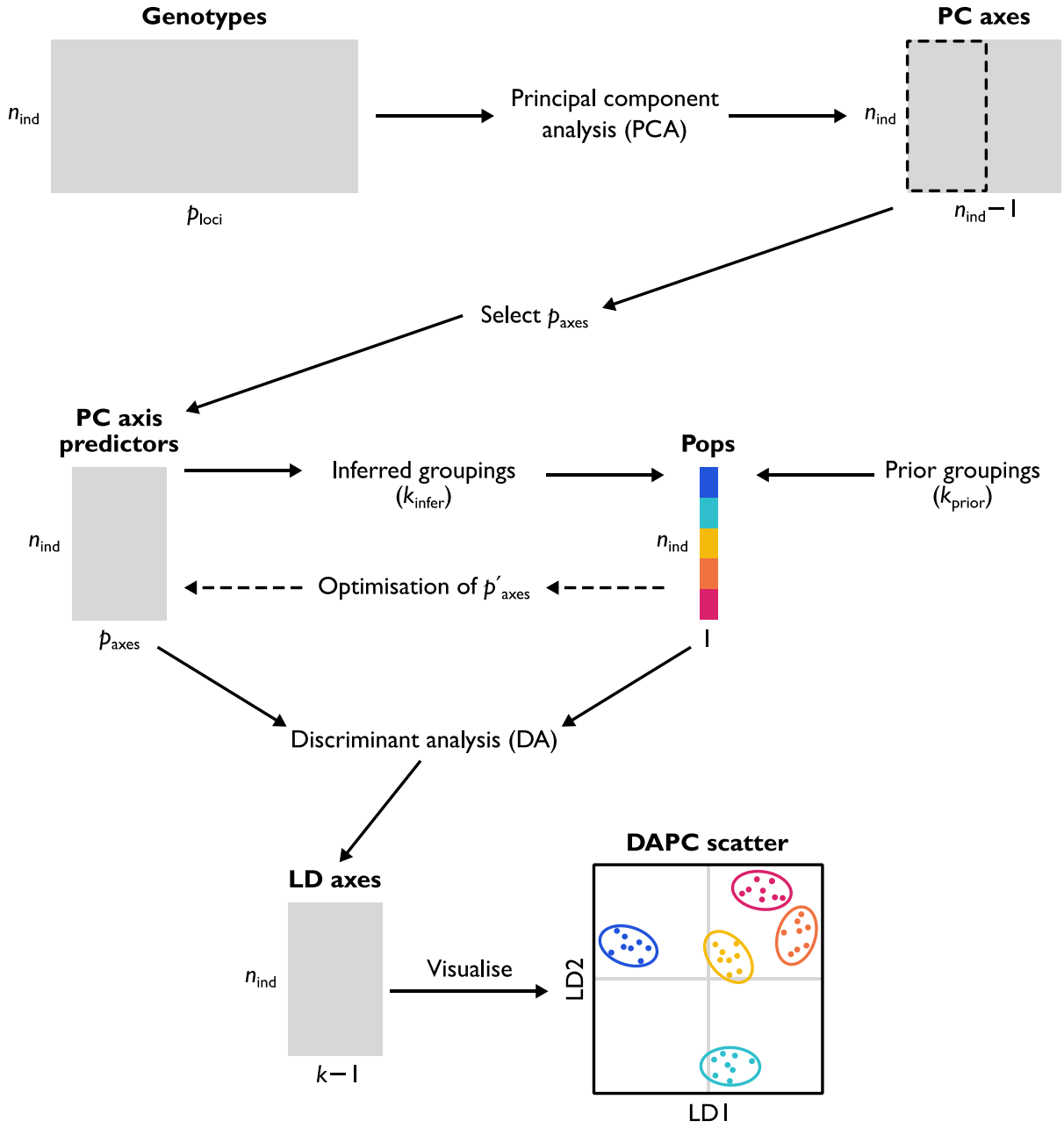

Figure 1. Schematic representation of a DAPC (discriminant analysis of principal components). A genotype matrix comprising n_{ind} samples genotyped at p_{loci} genetic loci is subjected to a PCA (principal component analysis). Assuming that $n_{ind} < p_{loci}$, there are $n_{ind} - 1$ PC axes with 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 populations. Assuming $p_{axes} \geq k - 1$, a set of $k - 1$ 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.

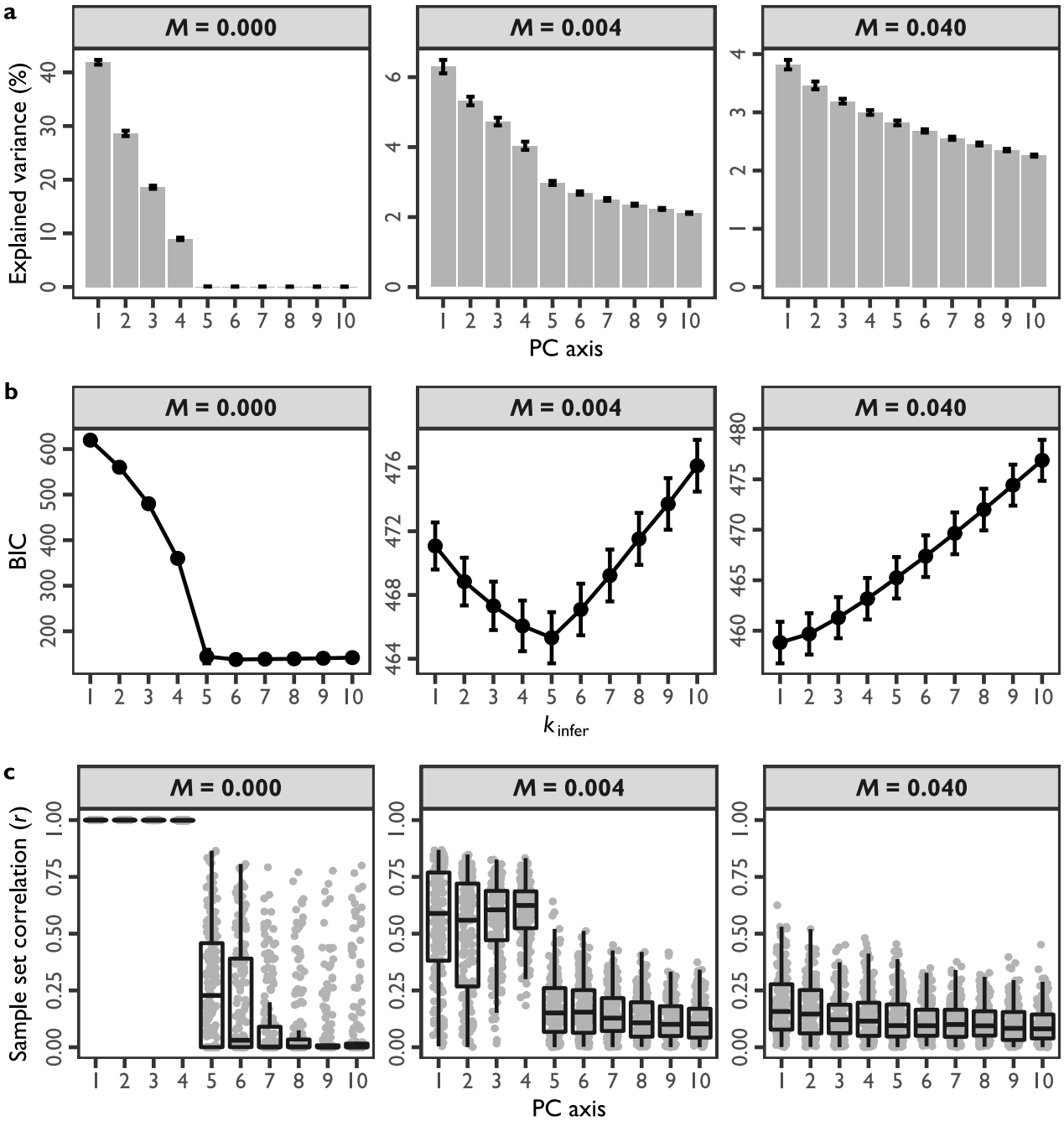

Figure 2. Principal component analysis under different migration scenarios. (a) Average explained variance for PC axes across simulated sample sets, using the *replicate simulation dataset*. The *x*-axis represents the first 12 PC axes. The *y*-axis is the amount of explained genotypic variance (percent of total) ascribed to each PC axis. Grey bars indicate the average across replicate simulations with the 95% confidence intervals. (b) The inferred number of populations by *K*-means clustering, using the *replicate simulation dataset*. The *x*-axis is the inferred number of clusters using p_{axes} = 80 PC axes (but see also Figure S4). 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. (c) 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 sample sets from the same simulated metapopulation. Boxplots summarise the distribution of correlations across sample set pairs.

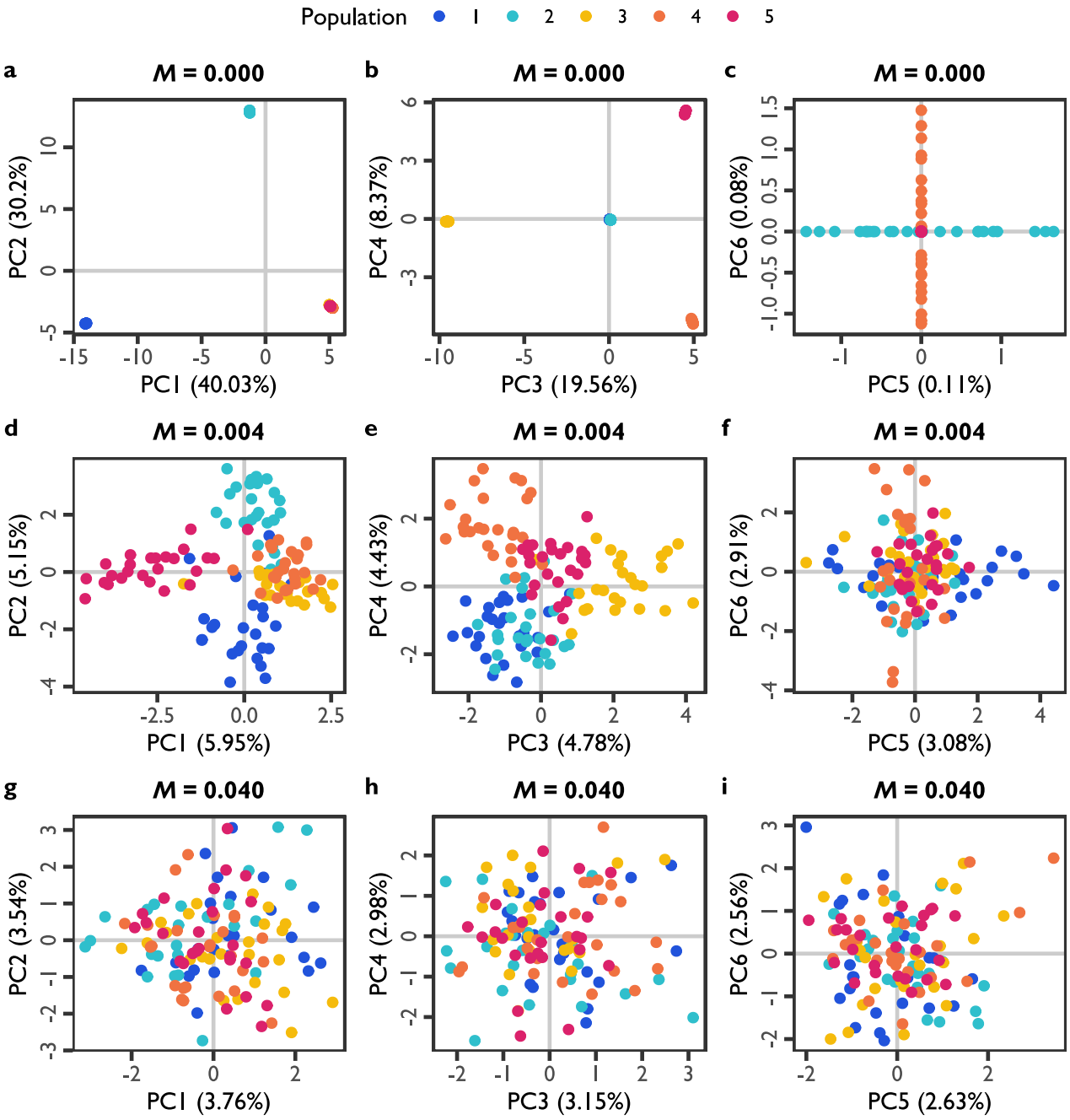

Figure 3. Exemplar scatterplots of individuals projected into PC space. The *x*-axis and *y*-axis, respectively, represent different pairs of PC axes. Numbers in parentheses indicate the amount of explained genotypic variance captured by each PC axis (percent of total). Each point represents an individual, 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.

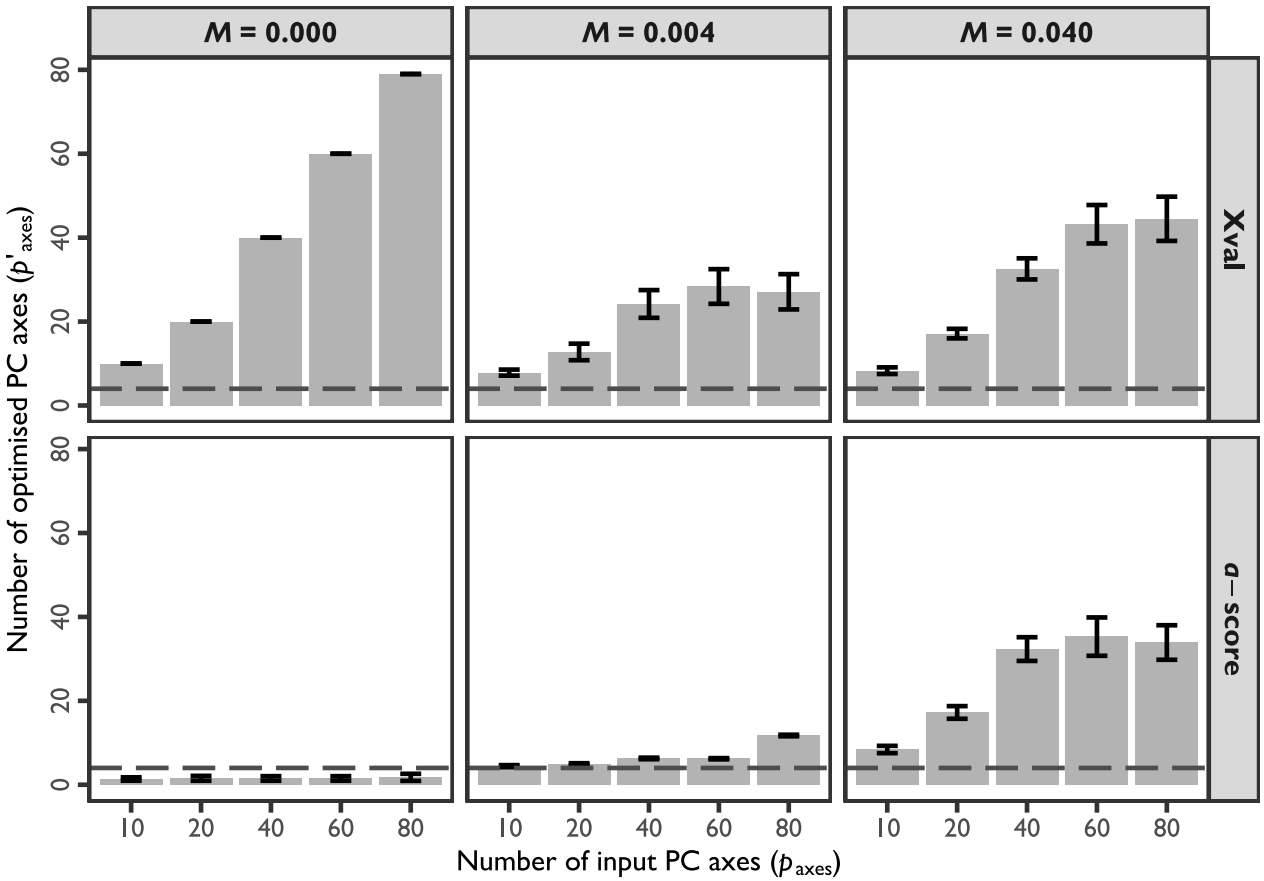

Figure 4. Optimisation of PC axes to fit DA models using the *replicate simulation dataset*. The *x*-axis is the number of PC axes used to initially fit the DA, p_{axes}. The *y*-axis is the number of optimised PC axes, p'_{axes}, using one of two optimisation algorithms available in the *adegenet* package. Panels represent combinations of optimisation algorithm (in rows) and migration scenario (in columns). Bars represent the mean p'_{axes} estimated across replicate simulations with 95% confidence intervals. The dashed lines demarcate $p'_{axes} = 4 = k - 1$.

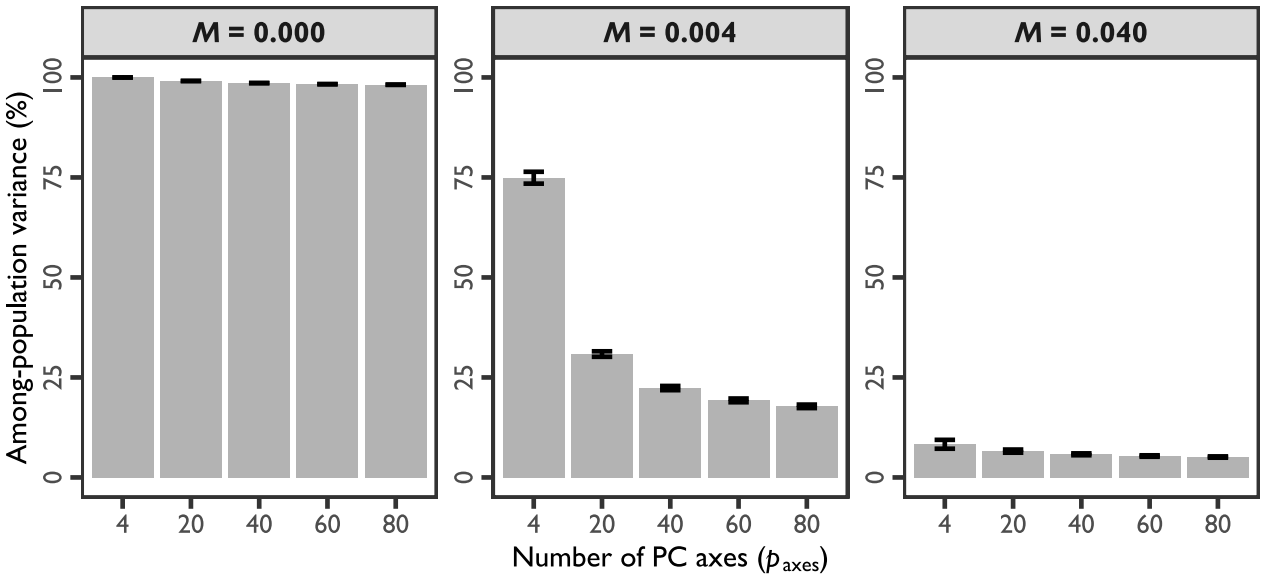

Figure 5. Average amount of among-population variance captured on PC axes using the *replicate simulation dataset*. The *x*-axis represents the total number of PC axes used in a DA. The *y*-axis represents the amount of variance (percent of total) that is among populations. Grey bars indicate the average across replicate simulations with 95% confidence intervals. Panels contain results for each migration scenario.

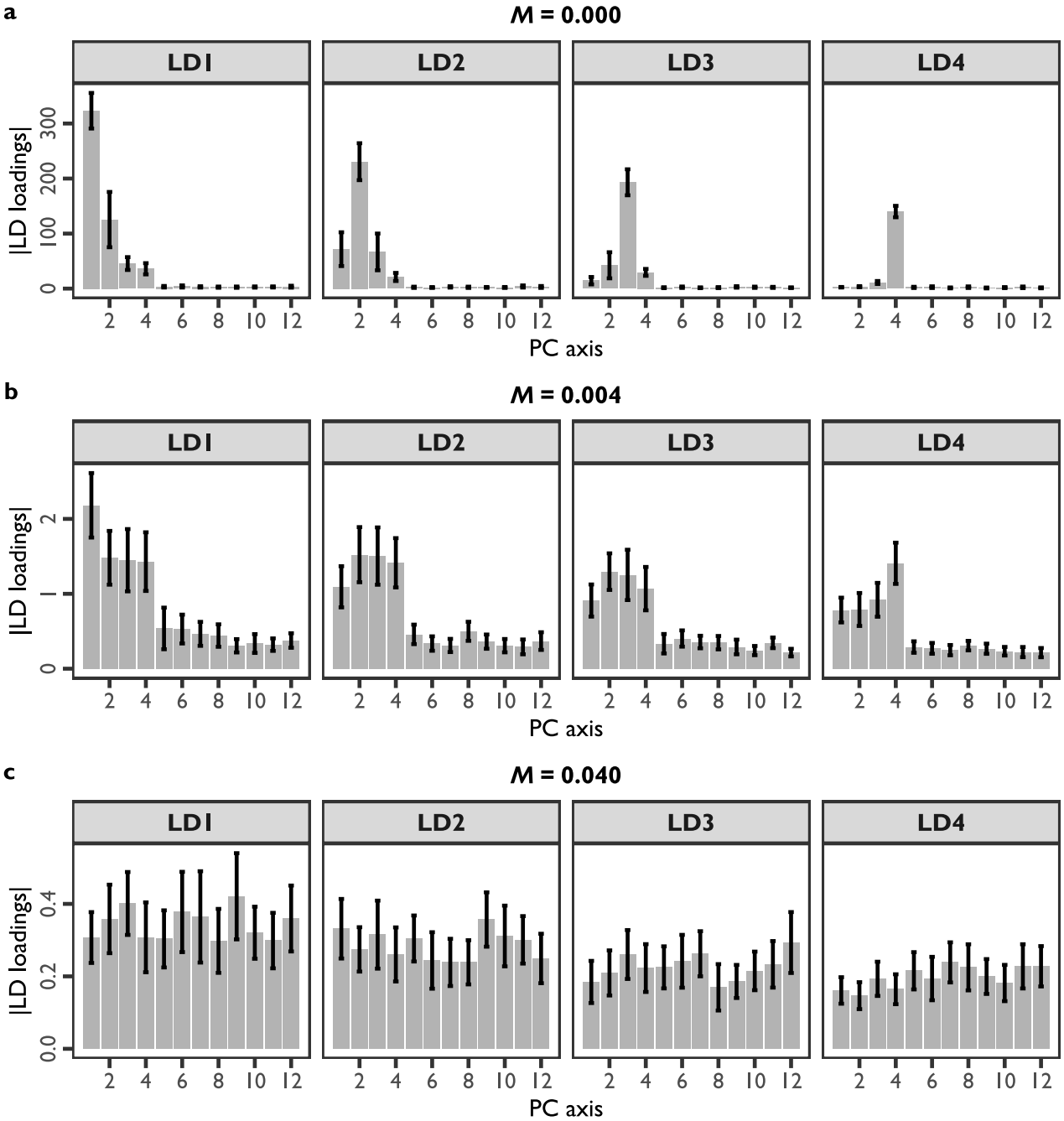

Figure 6. Average loadings of individual PC axes in DAs using the *replicate simulation dataset*. The *x*-axis represents the first 12 PC axes in DAs parameterised with p_{axes} = 80 PC axes. The *y*-axis represents the absolute loadings (contributions to variation) of PC axes onto discriminant functions 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.

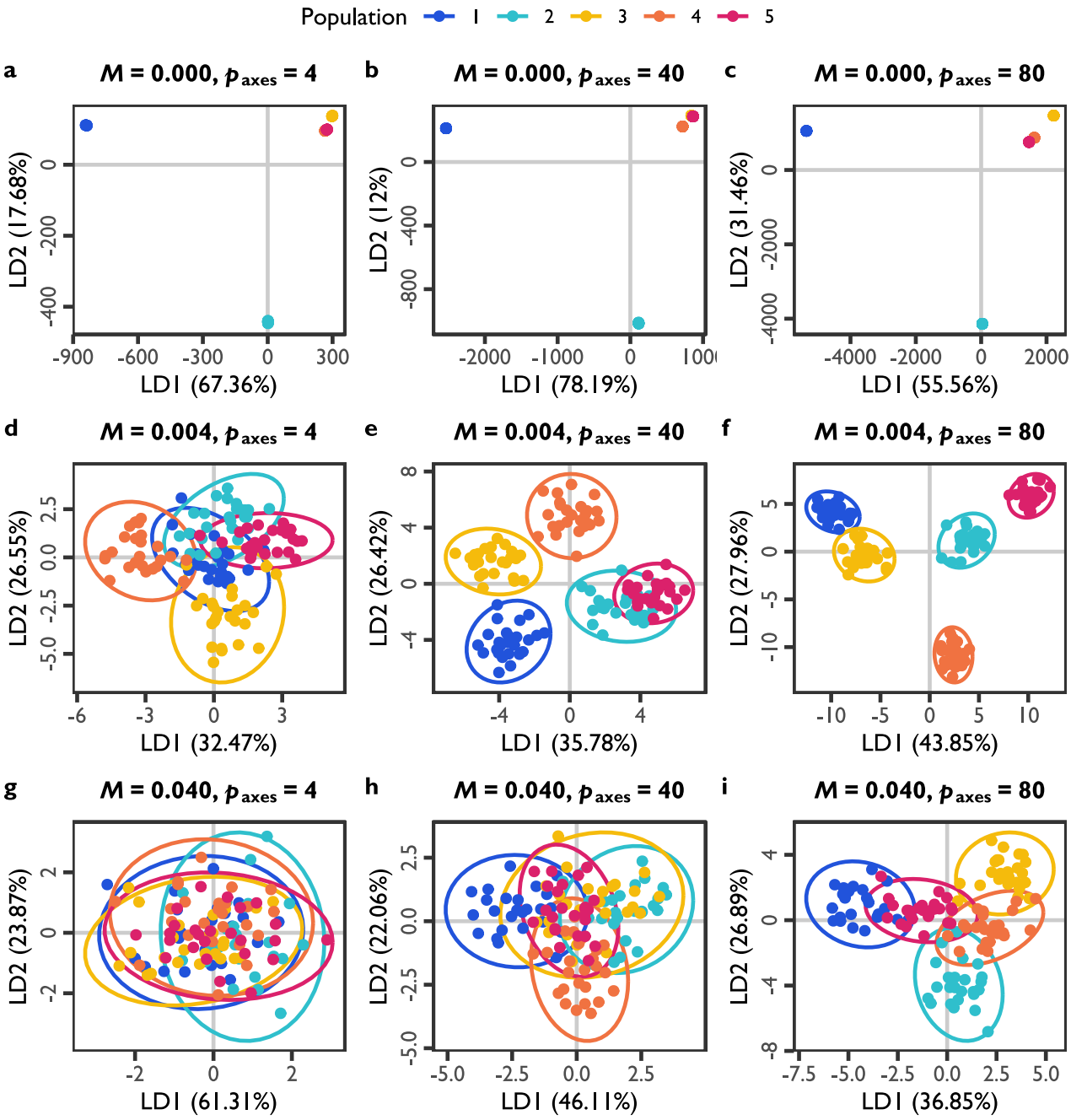

Figure 7. Exemplar scatterplots of individuals projected into LD space. The *x*-axis and *y*-axis represent the first and second LD axes, respectively, obtained from a DAPC parameterised with different numbers of PC axes, p_{axes}. Numbers in parentheses indicate the amount of explained among-population 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.

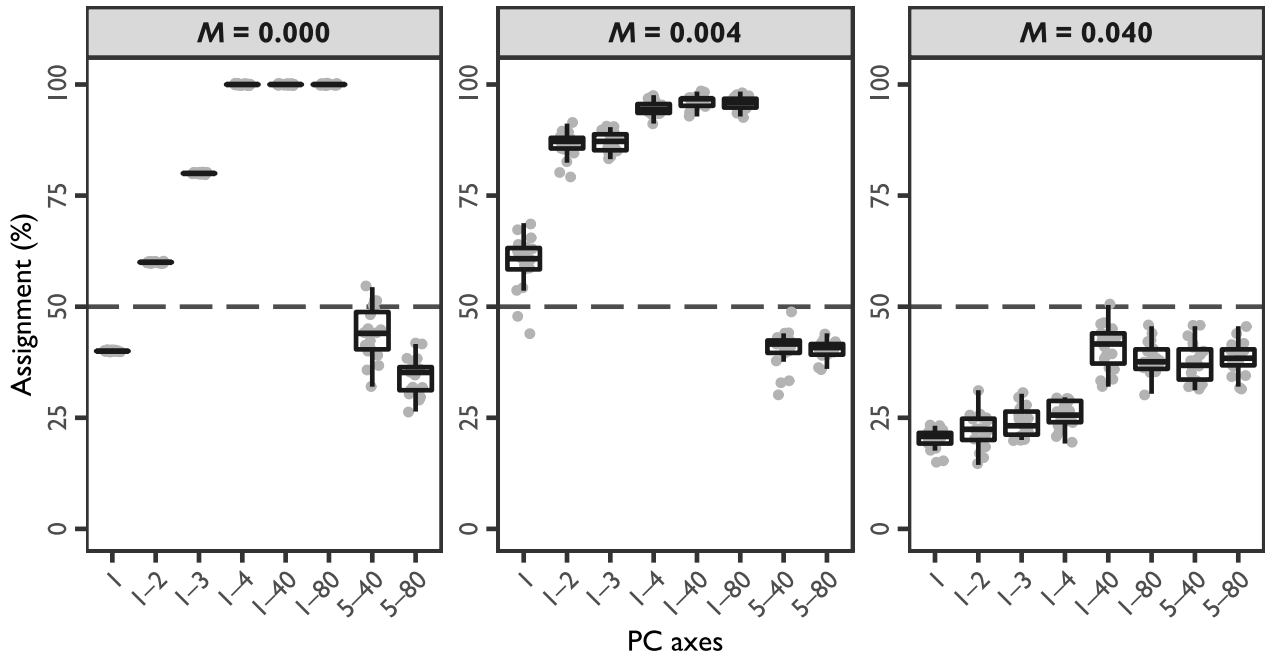

Figure 8. Correct assignment rates using DA models parameterised with different combinations of PC axes using the *singular simulation dataset*. DA models were constructed with a *training* sample set and used to assign populations to 19 *testing* sample sets. The *x*-axis represents the combinations 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) $M = 0.040$.

Supplementary Information: 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.

563 ## Population assignment

564 Population assignment analyses were conducted using the *singular simulation dataset*. One of the
565 independent sample sets was used as a *training* sample set, with the remaining 19 sample sets used as
566 *testing* sample sets. PCA of allelic covariances in the *training* sample set was conducted using the PRCOMP
567 function with the arguments "scale=FALSE, center=TRUE". A series of DAPC models were manually fit using
568 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–
569 80 PC axes. First, the LDA function from the *MASS* package was used to perform a DA using these different
570 combinations of PC axes. Second, individuals from each *testing* dataset were projected into the same PC
571 space as the *training* sample set, using eigenvectors from the *training* sample set PCA. Third, the DA model
572 constructed on the *training* sample set was used to project individuals from the *testing* sample sets into the
573 *training* sample set's LD space. Finally, population assignment was performed using R's PREDICT function.

574

Supplementary Information: Figures

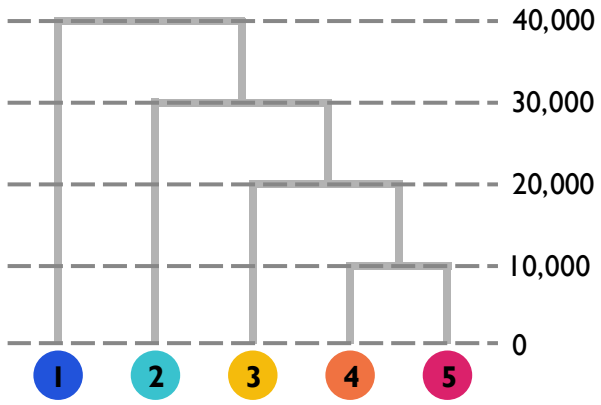

Figure S1. Simulated population divergence times. Each population (1–5) was comprised of 500 diploid individuals. Annotated nodes demarcate the number of generations in the past for each divergence event.

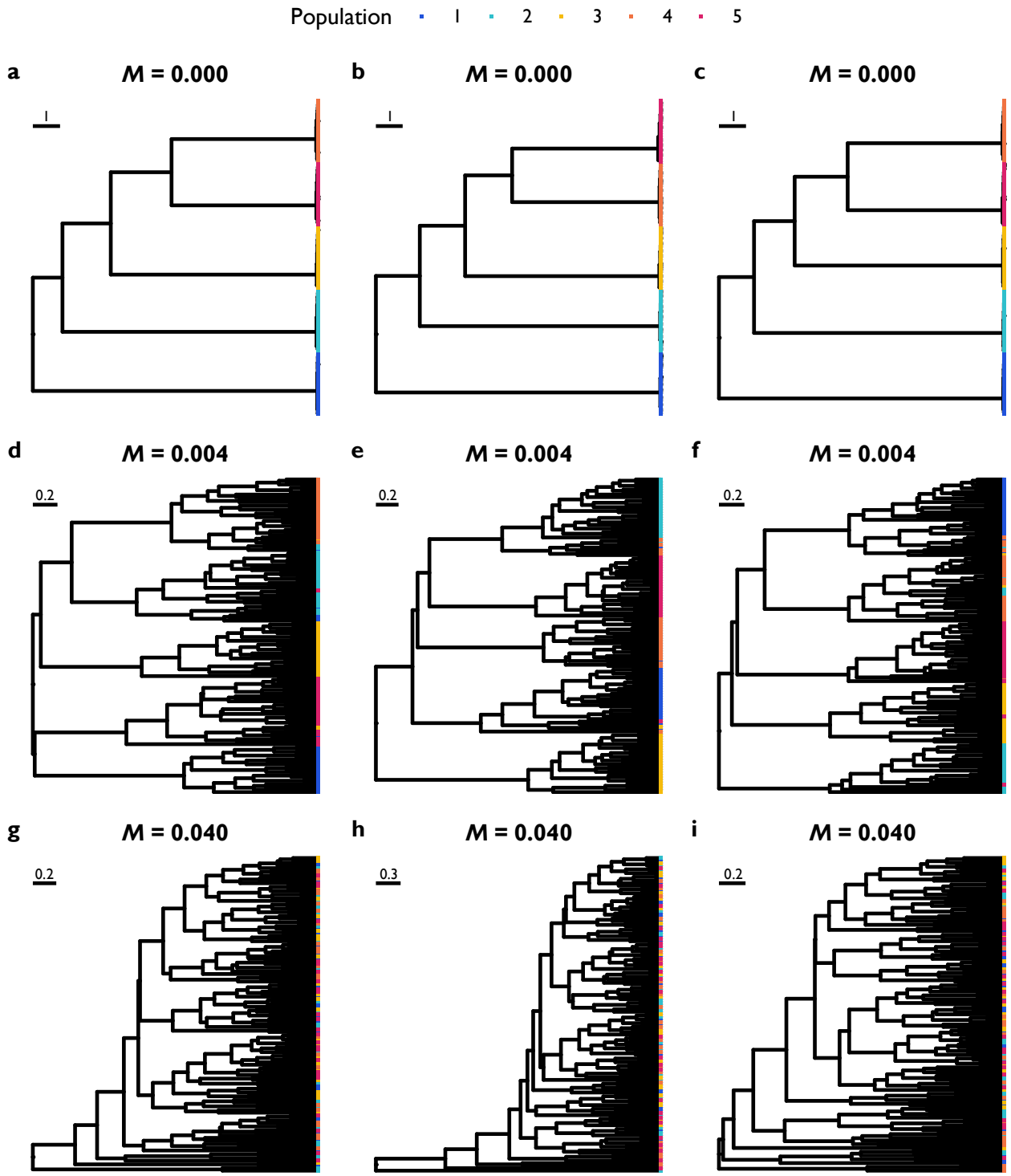

Figure S2. Hierarchical clustering of simulated samples in PC space from the *replicate simulation dataset*. Euclidean distance was estimated between all pairs of individuals using PC axes 1–4. Distance matrices were subjected to hierarchical clustering using the UPGMA method (unweighted pair group method 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.

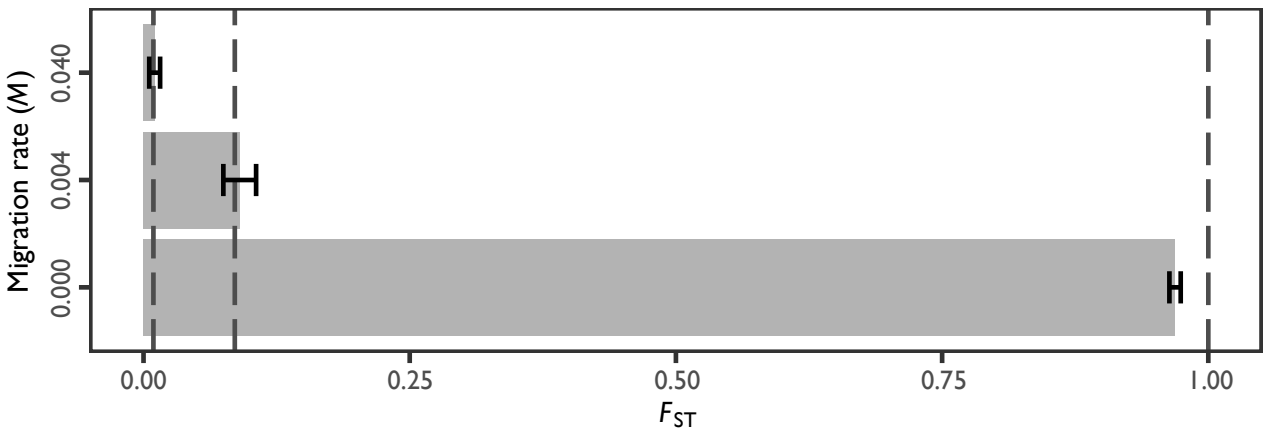

Figure S3. F_{ST} among populations in the *replicate simulation dataset*. The *y*-axis is the simulated migration rates and the *x*-axis is the estimated F_{ST}. Grey bars depict the mean F_{ST} across replicate simulations with 95% confidence intervals. Dashed lines demarcate the expected F_{ST} values 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).

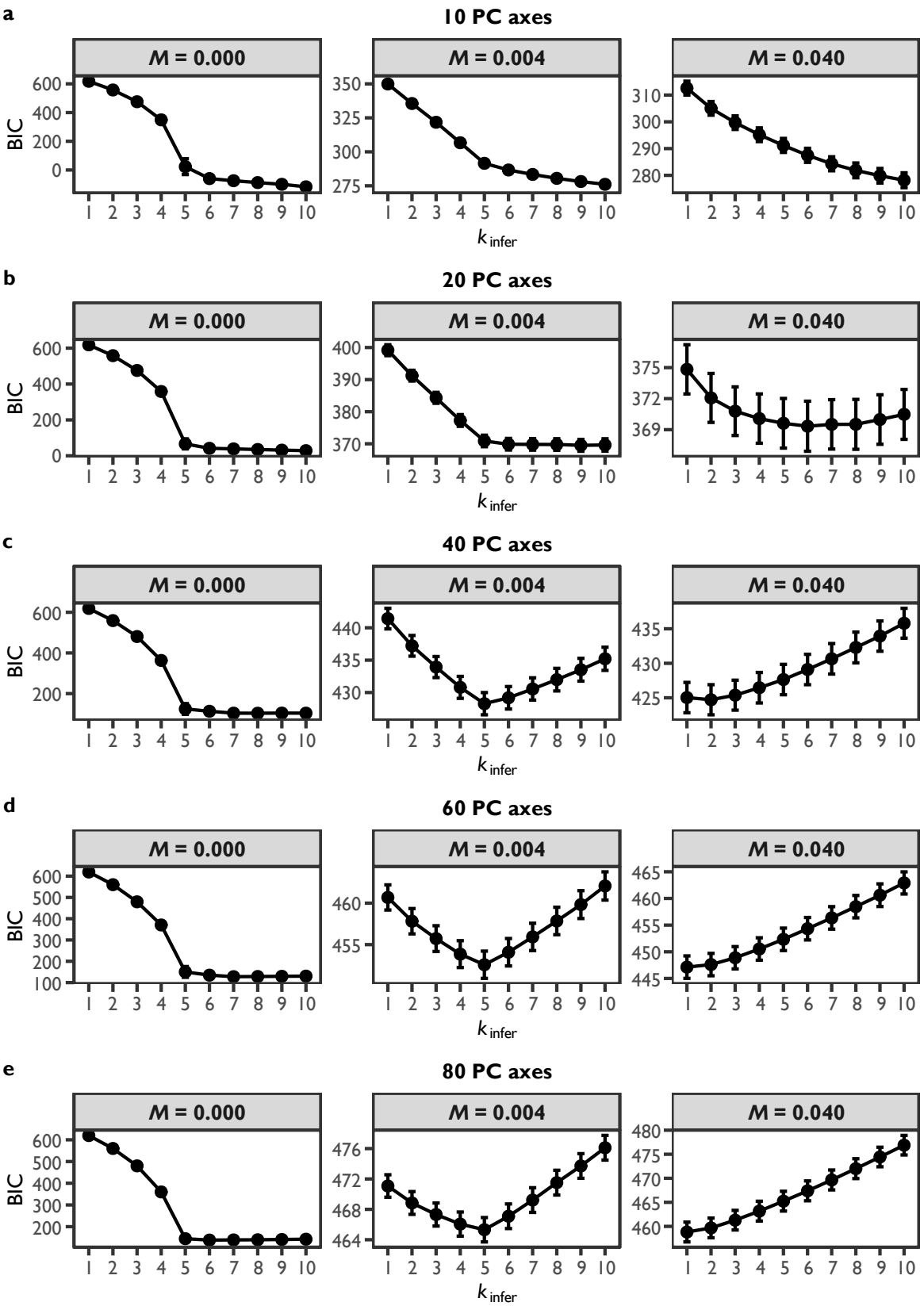

Figure S4. The inferred number of populations by *K*-means clustering, using the *replicate simulation dataset*. The *x*-axis is the inferred number of clusters. The *y*-axis is the associated BIC score (lower values are more likely). Points represent the mean BIC across replicate simulations with 95% confidence intervals. Panels represent different migration scenarios. Number of initial PC axes, p_{axes}: (a) 10, (b), 20, (c), 40, (d), 60, and (e) 80.