A Roadmap to Robust Discriminant Analysis of Principal 1 Components (DAPC)

Catherine Cullingham, Rhiannon M. Peery, Joshua M. Miller

A roadmap to robust Discriminant Analysis of Principal Components (DAPC)

Cullingham, CI[1], Peery RM[1] Miller JMM[2]

[1]Department of Biology, Carleton University, Ottawa, Ontario, Canada

[2]Biological Sciences, MacEwan University, Edmonton, Alberta, Canada

Cullingham ORCID: https://orcid.org/0000-0002-6715-0674

Peery ORCID: https://orcid.org/0000-0003-4711-4702

Miller ORCID: https://orcid.org/0000-0002-4019-7675

Corresponding author: Catherine Cullingham, catherine.cullingham@carleton.ca

Identification of population structure is a common goal for a variety of applications, including conservation, wildlife management, and medical genetics. The outcome of these analyses can have far reaching implications; therefore, it is important to ensure an understanding of the strengths and weaknesses of the methodologies used. Increasing in popularity, the Discriminant Analysis of Principal Components (DAPC) method incorporates combinations of genetic variables (alleles) into a model that differentiates individuals into genetic clusters. However, users may not have a full understanding of how to best parameterize the model. In this issue of Molecular Ecology Resources, Thia (2022) looks under the hood of the DAPC. Using simulated data, he demonstrates the importance of careful parameter selection in developing a DAPC model, what the implications are for over-fitting the model, and finally, how best to evaluate the results of DAPC models. This work highlights the issues that can arise when over-parameterizing the DAPC model when gene flow is high among clusters and provides important guidelines to ensure researchers are making conclusions that are biologically relevant.

The identification of clusters and the estimation of their divergence using genetic data has been a pillar of population genetics for over 70 years and the implications of these analyses can have far reaching impacts. For example, in the area of conservation biology, where resources are often limited, incorrect identification of significantly differentiated groups can result in either wasted efforts if groups are unnecessarily split, or inflated population size estimates if true population structure is not identified (Fallon, 2007). Thus, rigorous and accurate assessment of population structure is necessary (Haig et al., 2016). Similar to the maturing of the use of other clustering software (e.g. STRUCTURE, Pritchard, Stephens, & Donnelly, 2000)), as a community we need to establish a set of best practices for more recent software designed to find genetic structure.

The program STRUCTURE (Pritchard et al., 2000) has become the standard method for assessing population structure. Concurrent with its popularity, several papers have explored the limits and developed best practices for STRUCTURE (Gilbert et al., 2012; Janes et al., 2017; Lawson, van Dorp, & Falush, 2018). However, as the number of loci included in analyses increases, researchers often note the extensive time requirements as a barrier to its use (Wang, 2022). Given studies using datasets with thousands, to hundreds of thousand loci are now common, more expedient methods are needed. Discriminant analysis of principal components was introduced by Jombart, Devillard, & Balloux (2010) to address not only issues with analysis of large datasets, but also as a method that does not rely on the population genetic model assumptions associated with STRUCTURE. Discriminant analysis of principal components can be used to identify groups when they are unknown, visualize complex population structure, identify genomic regions driving population differences, and test assignment of individuals to clusters. Given this broad range of utility, together with the ease of use, it is now the third most cited method for assessing population structure (Figure 1).

Discriminant analysis of principal components involves a two-step process. The first is to use a Principal Component Analysis (PCA) to explain the population level variation among uncorrelated combinations of alleles by generating eigenvectors that summarize the covariance matrix generated from the data. The second is to select components that describe the between cluster variation and use those in a Discriminant Analysis (DA), which builds a model that can predict the population for each individual. The number of components selected from the PCA is critical as selecting too many axes will result in overfitting the model and create an inflated estimate of the amount of differentiation among clusters (Thia 2022). While the developers of

the DAPC method indicate the importance of this step, they do not provide a defined set of rules for selecting the components (Jombart & Collins, 2022). To address this issue, Thia (2022) examined simulated data for a five-population, finite-island model under three different levels of migration, low, medium, and high (panmixia). He analyzed the data using DAPC, while selecting different numbers of PC axes to use in the discriminant model. Based on these analyses, he highlights several important issues, and makes clear recommendations to ensure the proper use and interpretation of DAPC.

The selection of the number of PC axes has primarily relied on three approaches (Miller, Cullingham, & Peery, 2020): 1) selecting the number of PC axes which explain an arbitrary amount of variance in the dataset (often ≥ 80%), 2) the *xvalDapc* function, which uses a training and testing set to find the optimal trade-off between selecting too few, and too many PC axes, or 3) the *optim.a.score* function, which looks at the assignment success of individuals based on different numbers of PC axes retained, and selects the number of PCs that maximizes assignment success. The work by Thia (2022) highlights caveats to consider for each of these methods. The first method often leads to inclusion of components which explain individual differences, rather than just including those which capture between population differences. When there is no migration (F_{ST} = 0.99 in simulated data), there is little impact on the conclusions made because the majority of variance is between populations. However, when migration is medium to high (F_{ST} = 0.09 & 0.0009, respectively) there is an inflation of how differentiated populations appear (Figure 8 in Thia 2022). The second method suffered from similar issues to the first, where the *xvalDapc* function indicates more PC axes be selected than those that just explained population structure in the simulated datasets, again resulting in an over-fit model. The third method was the most conservative, but again returned inflated values of PC axes to obtain when migration was medium to high. This is highly relevant given a review in 2020 found that the majority of studies using DAPC do not report the number of PC axes retained (64%) and for those that do, the range of reported values (2-600, mean = 57) suggests many studies are selecting far too many axes (Miller et al., 2020).

As an alternative to these approaches, Thia (2022) suggests selecting the number of PC axes using the *K*-1 rule from Patterson, Price, & Reich, (2006), where *K* is the number of genetic clusters. This initially seems like a very conservative approach; however, Thia (2022) clearly demonstrates these components capture the between population differences, and when migration is moderate, provide the best model for predicting population membership. Including additional PC axes can inflate or obscure the true population structure.

Of course, many studies do not necessarily know the number of populations represented by their samples, and therefore may be interested in using DAPC to determine the most likely value of *K*. Thia (2022) recommends using alternative programs for finding *K*, however, as noted in Miller et al. (2020) most researchers are using more than one method to corroborate a choice of *K* and several authors use DAPC along with STRUCTURE or other admixture programs. To facilitate decision making when employing DAPC we have generated a decision tree that includes some recommendations made in Thia (2022) together with best practices in population genomic analyses that are tailored to the overall goal of the study (Figure 2). In some instances, conducting a PCA without a DA may be the most appropriate approach to use. As noted by Thia (2022) and in Miller et al., (2020) reporting of decisions and parameters are critical for evaluation of the DAPC analyses itself and for reproducibility.

In recognizing the important utility of a PCA based method, we must also mention the Achilles heel of the PCA – no method is robust to poorly curated datasets. The matrix of alleles that is used for the PCA needs to be properly filtered and pruned before analysis. Errors and bias in the data can strongly influence the overall structure of the PC plot (e.g., Patterson et al., 2006), especially when genetic differentiation is low. Missing data, a common feature of the genotyping-by-sequencing method, will also influence the PC plot, sometimes suggesting population structure where none exists (e.g., Yi & Latch, 2022). No matter the study objective, best practices for assessing data integrity in genomic datasets need to be followed prior to examining population structure (Laurie et al., 2010; Shafer et al., 2017).

Identifying and characterizing population structure will continue to be an important goal in molecular ecology. To this end, methods that are rapid, consistent, and can define clusters when genetic differentiation is limited are desirable. Discriminant analysis of principal components fits these criteria, yet the parameterization of the model has not been thoroughly addressed in the literature. Thia (2022) address some of these outstanding concerns, and demonstrates it is a powerful tool, but can lead to false conclusions when too many PC axes are included in the model, especially when genetic differentiation is moderate to low. Thia (2022) then goes on to present a list of recommendations (included in Figure 2), that if followed, will allow researchers to arrive at conclusions that are relevant for the biology of their species. We expect with these clear recommendations, the use of this method will continue to increase, and users will benefit from the utility of the tool.

Data accessibility

Data generated from the Web of Science search and the R code for generating Figure 1 can be found at https://doi.org/10.5683/SP3/UPONUS.

References

Fallon, S. M. (2007). Genetic Data and the Listing of Species Under the U.S. Endangered Species Act. *Conservation Biology*, *21*(5), 1186–1195.

Gilbert, K. J., Andrew, R. L., Bock, D. G., Franklin, M. T., Kane, N. C., Moore, J.-S., … Vines, T. H. (2012). Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. *Molecular Ecology*, *21*(20), 4925–4930.

Haig, S. M., Miller, M. P., Bellinger, R., Draheim, H. M., Mercer, D. M., & Mullins, T. D. (2016). The conservation genetics juggling act: integrating genetics and ecology, science and policy. *Evolutionary Applications*, *9*(1), 181–195.

Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., Cullingham, C. I., & Andrew, R. L. (2017). The K = 2 conundrum. *Molecular Ecology*.

Jombart, T., & Collins, C. (2022). A tutorial for Discriminant Analysis of Pricipal Components (DAPC) using *adegenet 2.1.6* (p. 43). p. 43.

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),

94.

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., … Weir, B. S. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591–602.

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1).

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 2020 125:5*, *125*(5), 269–280.

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genetics*, *2*(12), e190.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907–917.

Thia, J. (2022). Guidelines for standardising the application of discriminant analysis of pricipal components to genotype data. *Molecular Ecology Resources*, *XX*, XXX–XXX.

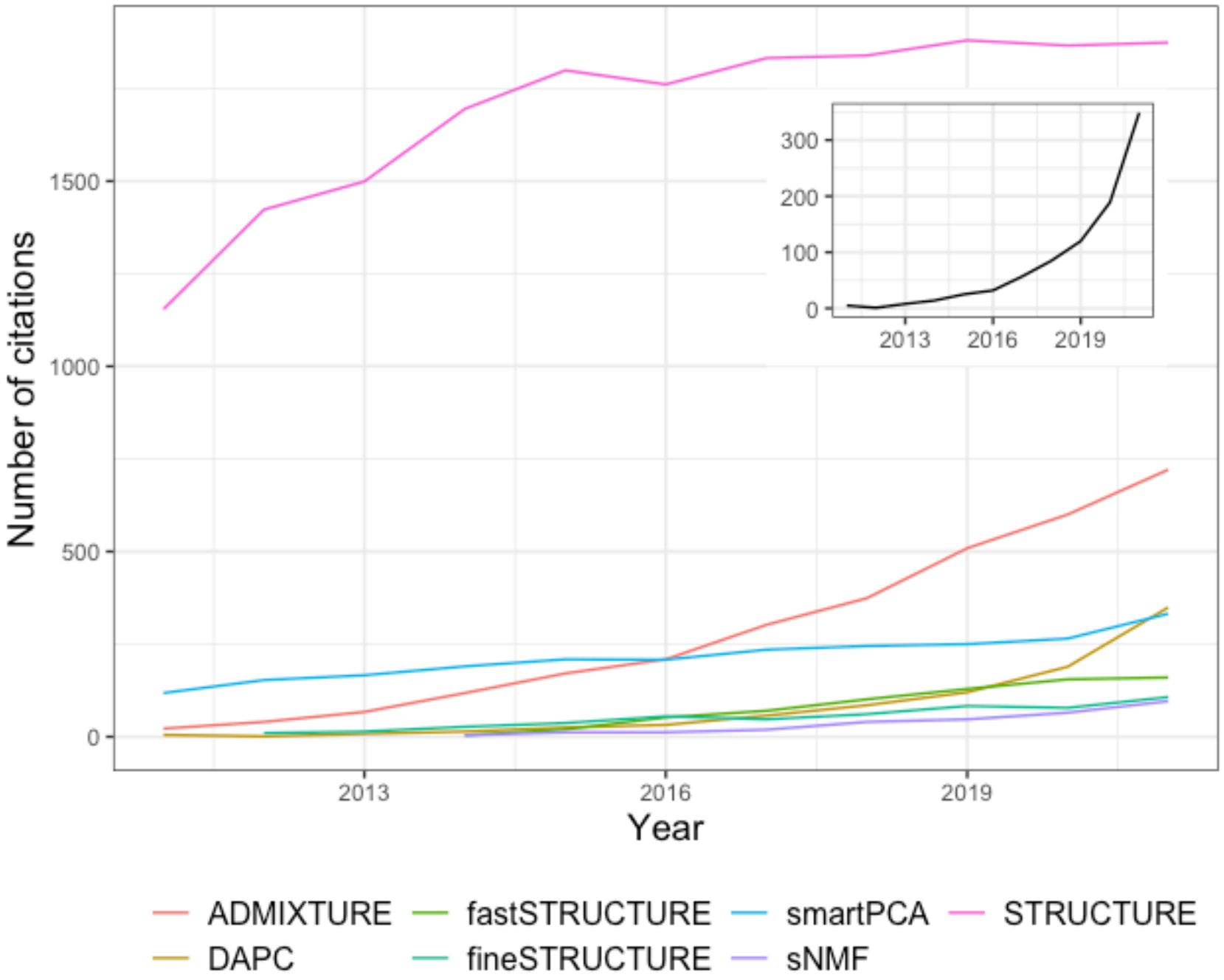
Wang, J. (2022). Fast and accurate population admixture inference from genotype data from a few microsatellites to millions of SNPs. *Heredity 2022 129:2*, *129*(2), 79–92.

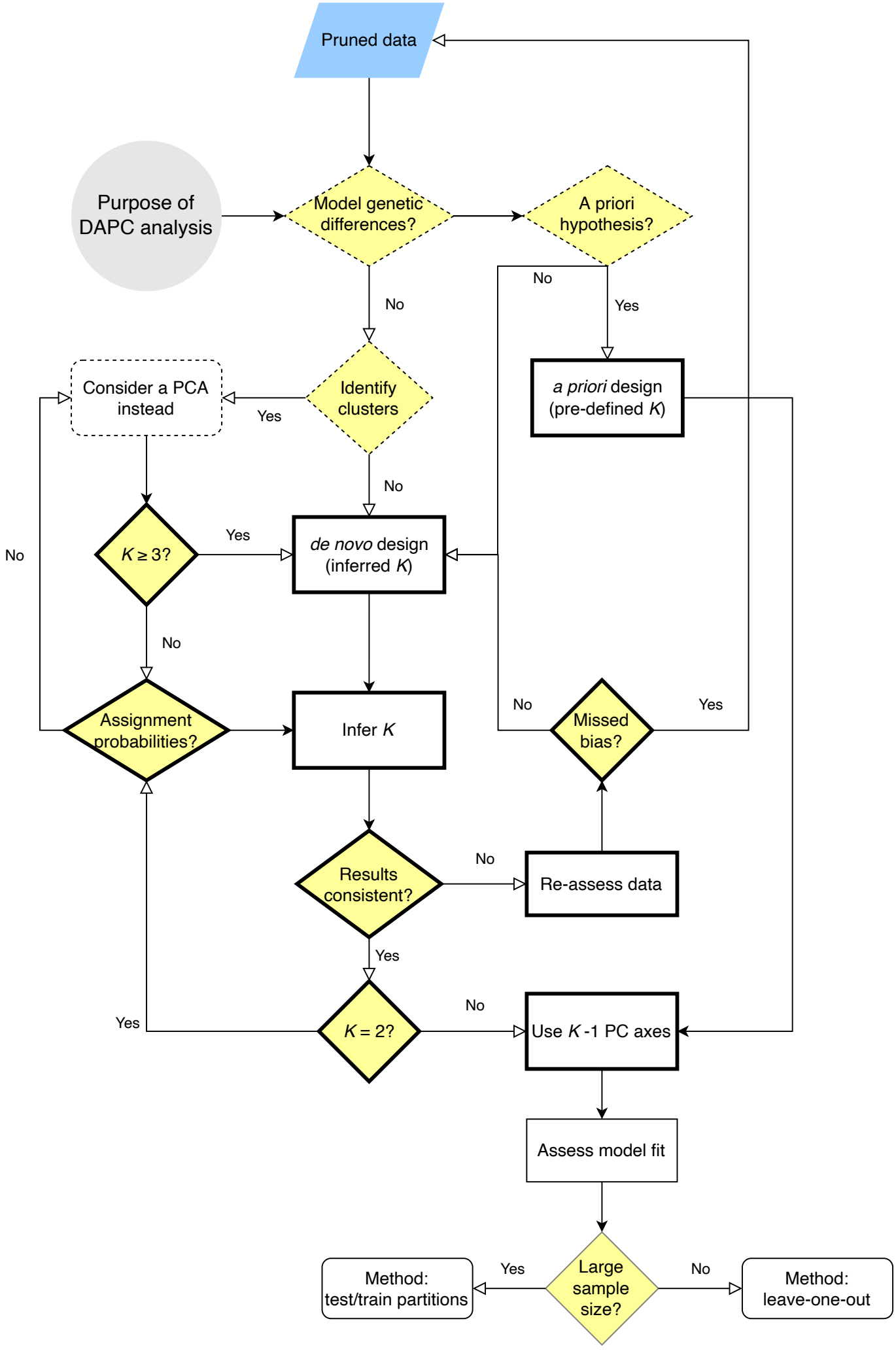
Yi, X., & Latch, E. K. (2022). Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. *Molecular Ecology Resources*, *22*(2), 602–611.

Figure captions

Figure 1. Total number of citations, calculated via reverse citation search of each program's initial publication in Web-of-Science on Sept. 9, 2022, for popular clustering algorithms. The popularity of DAPC has steadily increased since its publication in 2011 (inset).

Figure 2. Decision tree for DAPC analyses designed in draw.io. The first phase is goal setting (shapes with stippled outlines), the second phase determines the model parameters and user inputs (bold outlines), and the final phase tests the model fit (thin, solid outline). The data pruning and reporting phases recommended by (Laurie et al., 2010; Miller et al., 2020; Shafer et al., 2017) are not shown.


Number of citations
1500
1000
500
0
2013
2016
2019
Year
300
200
100
0
2013
2016
2019
ADMIXTURE
DAPC
fastSTRUCTURE
fineSTRUCTURE
smartPCA
sNMF
STRUCTURE


Pruned data
Purpose of DAPC analysis
Model genetic differences?
A priori hypothesis?
No
Yes
No
Yes
Consider a PCA instead
Identify clusters
Yes
a priori design (pre-defined K)
No
K ≥ 3?
Yes
de novo design (inferred K)
No
No
Assignment probabilities?
Infer K
No
Missed bias?
Yes
Results consistent?
No
Re-assess data
Yes
Yes
K = 2?
No
Use K -1 PC axes
Assess model fit
Large sample size?
Yes
Method: test/train partitions
No
Method: leave-one-out