Received: 16 June 2017 | Accepted: 6 September 2017

DOI: 10.1111/2041-210X.12897

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

assignPOP: An R package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework

Kuan-Yu Chen[1] | Elizabeth A. Marschall[1] | Michael G. Sovic[2,3] | Anthony C. Fries[2,4] | H. Lisle Gibbs[2] | Stuart A. Ludsin[1]

[1]Aquatic Ecology Laboratory, Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH, USA

[2]Department of Evolution, Ecology and Organismal Biology and Ohio Biodiversity Conservation Partnership, The Ohio State University, Columbus, OH, USA

[3]Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA

[4]United States Air Force School of Aerospace Medicine, Wright-Patterson AFB, OH, USA

Correspondence
Kuan-Yu Chen
Email: chen.1735@osu.edu

Handling Editor: Timothée Poisot

Abstract

1. The use of biomarkers (e.g., genetic, microchemical and morphometric characteristics) to discriminate among and assign individuals to a population can benefit species conservation and management by facilitating our ability to understand population structure and demography.
2. Tools that can evaluate the reliability of large genomic datasets for population discrimination and assignment, as well as allow their integration with non-genetic markers for the same purpose, are lacking. Our R package, *assignPOP*, provides both functions in a supervised machine-learning framework.
3. *assignPOP* uses Monte-Carlo and *K*-fold cross-validation procedures, as well as principal component analysis, to estimate assignment accuracy and membership probabilities, using training (i.e., baseline source population) and test (i.e., validation) datasets that are independent. A user then can build a specified predictive model based on the relative sizes of these datasets and classification functions, including linear discriminant analysis, support vector machine, naïve Bayes, decision tree and random forest.
4. *assignPOP* can benefit any researcher who seeks to use genetic or non-genetic data to infer population structure and membership of individuals. *assignPOP* is a freely available R package under the GPL license, and can be downloaded from CRAN or at https://github.com/alexkychen/assignPOP. A comprehensive tutorial can also be found at https://alexkychen.github.io/assignPOP/.

KEYWORDS
assignment analysis, machine learning, population classification, quantitative genomics

1 | INTRODUCTION

The ability to discriminate among resident and immigrant individuals in local populations and then identify the source populations from which immigrants originated can facilitate species conservation and management (Manel, Gaggiotti, & Waples, 2005). Numerous types of data have been applied towards this goal (reviews in Begg & Waldman, 1999; Cadrin, Kerr, & Mariani, 2013; Hobson, 1999; Waples & Gaggiotti, 2006). Most prominent across all taxa has been the use of genetic markers (e.g., microsatellites, SNPs); however, other data types have also been used, including artificial tags and various types of natural tags such as morphological traits (Cadrin, 2000), parasites (Mackenzie,

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2017 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

2002), fatty acid composition (Czesny, Dabrowski, Christensen, Van Eenennaam, & Doroshov, 2000), and elemental or isotopic composition of fish otoliths (Campana, Fowler, & Jones, 1994), bird feathers (Clegg, Kelly, Kimura, & Smith, 2003) and invertebrate shells (Becker, Levin, Fodrie, & McMillan, 2007). While the use of genetic markers has improved our ability to understand population-level phenomena, including life-history variation (Gibbs et al., 1990; Koehler, Pearce, Flint, Franson, & Ip, 2008; Ross, 2001) and population dynamics (Hardesty, Hubbell, & Bermingham, 2006; Hellberg, Burton, Neigel, & Palumbi, 2002), current methods used for integrating genetic data with information from other markers are limited and impede our ability to fully understand population structure and demography.

One major methodological limitation is the inability to robustly evaluate the power of a given dataset to discriminate among local populations. Two main issues limit our ability to achieve this goal: (1) test data that are not independent from the data used to develop the classification functions from source populations (i.e., training data or baseline data; Anderson, 2010; Waples, 2010) and (2) unbalanced sample sizes among the source (baseline) populations, which inherently lead to assigning more individuals to the source population with the larger sample size (Wang, 2016). For example, the program GENECLASS2 (Piry et al., 2004), which uses jackknifing (leave-one-out) as its default method to assign individuals to the source populations, can upwardly bias assignment accuracy because the training data used in the validations are nearly identical (i.e., only differing by 1 data point, James, Witten, Hastie, & Tibshirani, 2013). Another widely used program, STRUCTURE (Pritchard, Stephens, & Donnelly, 2000), which uses a Bayesian approach to assign individuals to populations, also has been shown to provide erroneous results, if population sample sizes are unbalanced (Puechmaille, 2016; Wang, 2016).

The other major limitation is the inability to integrate different marker types and efficiently process the resulting high-dimensional data. Several studies have described the value of integrating different markers to better understand population structure (e.g., Bradbury, Campana, & Bentzen, 2008; Chabot, Hobson, Wilgenburg, McQuat, & Lougheed, 2012; Gómez-Díaz & González-Solís, 2007; Guillot, Renaud, Ledevin, Michaux, & Claude, 2012; Kelly, Ruegg, & Smith, 2005; Perrier et al., 2011; Smith & Campana, 2010). However, few methods exist for combining the various marker datasets into a single predictive model that can be used to classify individuals of unknown origin. As far as we are aware, only a small number of ecological studies (Perrier et al., 2011; Ruegg et al., 2016; Rundel et al., 2013; Smith & Campana, 2010) have attempted to ascertain the source origin of individuals using integrated genetic (e.g., microsatellites) and non-genetic (e.g., microchemistry) data. Unfortunately, the methods developed using a Bayesian framework to integrate data are not always reliable. For example, Smith and Campana (2010) found that the results of assignment success generated from integrated data were worse than when each data type was used independently. Thus, methods that can allow for successful integration of genetic and non-genetic marker data can still be improved.

Moreover, as genome-wide sequencing data become more easily obtained, methods that can reduce the dimensionality of large datasets (e.g., >10,000 SNPs) and allow their integration with other marker types are becoming increasingly important, especially when access to a high-capacity computing cluster is lacking. While the R package ADEGENET (Jombart, 2008; Jombart & Ahmed, 2011) does use discriminant analysis of principal components to allow for data dimensionality reduction in large genomic datasets (Jombart, Devillard, & Balloux, 2010), it does not allow one to easily integrate genetic and non-genetic data or allow for the separation of PCs between data types. These limitations point to the need for a versatile toolkit for discriminating between populations, which can integrate different data types, overcome issues associated with datasets being non-independent and unbalanced, and offer multiple classification methods.

The R package, *assignPOP*, which we describe below, fulfils these needs and ultimately can benefit researchers interested in understanding population structure and demographics. In brief, our package offers many novel features, including (1) a function to concatenate genetic and non-genetic data, (2) principal component analysis (PCA) to reduce data dimensionality while allowing genetic and non-genetic PCs to be separated, (3) resampling cross-validation to estimate assignment accuracy and membership probability, and (4) several machine-learning classification algorithms to build tunable predictive models. Built-in options also allow users to easily manipulate the sample size of training datasets, so that biases associated with unbalanced population sizes (Wang, 2016) and self-assignment components (Anderson, 2010; Waples, 2010) can be assessed and avoided. More details about the package are described below.

2 | DESCRIPTION

2.1 | Overview of analytical framework

To accurately ascertain population membership of a sample of individuals from a "mixed" population, the baseline data from the source populations that are used to develop classification (i.e., predictive assignment) functions should be free from bias. In *assignPOP*, these baseline data can consist of only genetic data, only non-genetic data, or a combination of genetic and non-genetic marker information. This information can consist of "features" that are biological or chemical in nature. By repeatedly creating new classification functions and subsequently performing assignment tests, using randomly sampled datasets of varying size, our resampling cross-validation procedures allow for the unbiased creation of training datasets that are "balanced" (i.e., of equal sample size) among source populations, as well as an evaluation of their predictive accuracy (Figure 1).

Below, we provide a brief overview of the analytical protocol that *assignPOP* follows (all steps below refer to Figure 1), followed by a demonstration of some of the many options offered in the package. Individuals are first divided into training and test groups, with sample sizes defined by the user (Step 1). Next, for each training dataset, one or multiple user-specified subsets of training features (i.e., genotypes and/or non-molecular data) are reduced in dimensionality using PCA (Step 2), the output of which are used to build a user-chosen machine-learning classification function (Step 3). The classification functions then are used to assign "test" individuals to a source population (Step 4). This entire process can be automatically repeated as many times

as the user specifies, using one of two user-chosen resampling cross-validation procedures (Step 5). It is this resampling, and the subsequent analysis of the descriptive statistics and data visualizations provided, which allows for an assessment of the reliability of the baseline data to accurately assign individuals to a source population. If the results of this evaluation are deemed satisfactory, then one can use the entire set of baseline data to perform a final assignment test on any unknown individuals from a mixed population (Step 6).

2.2 | Dataset splitting and resampling

Our *assignPOP* package offers two resampling, cross-validation procedures, Monte-Carlo (Xu & Liang, 2001) and *K*-fold (Rodriguez, Perez, & Lozano, 2010), which are used to initially split the data into training and test groups (Step 1) and then test the predictive accuracy of the training data (Step 5). When performing Monte-Carlo cross-validation (function *assign.MC*), *assignPOP* allows users to determine a set of proportions (or fixed numbers) of individuals from each source population to be used in the training dataset, with the remaining individuals being allocated to the test dataset. By allowing for the creation of randomly selected, independent training and test datasets that vary in their relative size, but that are still balanced in size among all of the source populations, the user can use the descriptive statistics and visualizations provided to assess how training (and test) sample size bias assignment results. In this way, our resampling approach avoids the limitations of other programs used for population assignment, such as the use of non-independent training and test datasets (sensu Anderson, 2010) and biased inference due to unequal sample size among source populations (Wang, 2016).

Because the Monte-Carlo procedure samples random individuals each time, and hence does not guarantee that every individual is sampled, we included a *K*-fold cross-validation (function *assign.kfold*) option as an alternative resampling procedure to help estimate membership probability across individuals. When using the *K*-fold method, individuals from each population are randomly divided into *K* groups. One group from each population is used as test individuals and the remaining $K - 1$ groups are used as training individuals to build the predictive model. Then assignment tests are performed until every group is tested, resulting in *K* tests. In this way, test individuals are independent from the training individuals, and every individual is guaranteed to be tested once. In the function *assign.kfold*, multiple *K*-values can be specified in a single analysis.

2.3 | Data importation and integration

To analyse genetic data, our package provides the function *read.Genepop* that allows users to import a GENEPOP format file (Rousset, 2008) into R. The function converts genetic data to a numeric format, following the method used in ADEGENET (Jombart, 2008). In this way, each allele is encoded as 0 (absent), 0.5 (heterozygote) or 1 (homozygote). To analyse non-genetic data, users can save their tabular data in a text or comma-delimited file and import it using the R basic functions *read.table* or *read.csv*. In addition, users can use the function *compile.data* to integrate genetic and non-genetic data. Categorical data (e.g., colours) in non-genetic data are converted to dummy variables for further analyses. Continuous data can be standardized ($M = 0$ and $SD = 1$), if the ranges of non-genetic data vary.

2.4 | Dimensionality reduction

We use PCA to reduce dimensionality in the training dataset, which is useful when there are a large number of markers. The PCA is always conducted on molecular data and is optional for the non-molecular data. The loadings of selected PCs are recalculated based on the training data of each independent iteration and used to calculate the scores (coordinates) for both training and test individuals. This PCA is internally performed when running the function *assign.MC* or *assign.kfold*, and by default it retains any PC that has an eigenvalue greater than 1 (the Kaiser-Guttman criterion; Guttman, 1954; Kaiser, 1960).

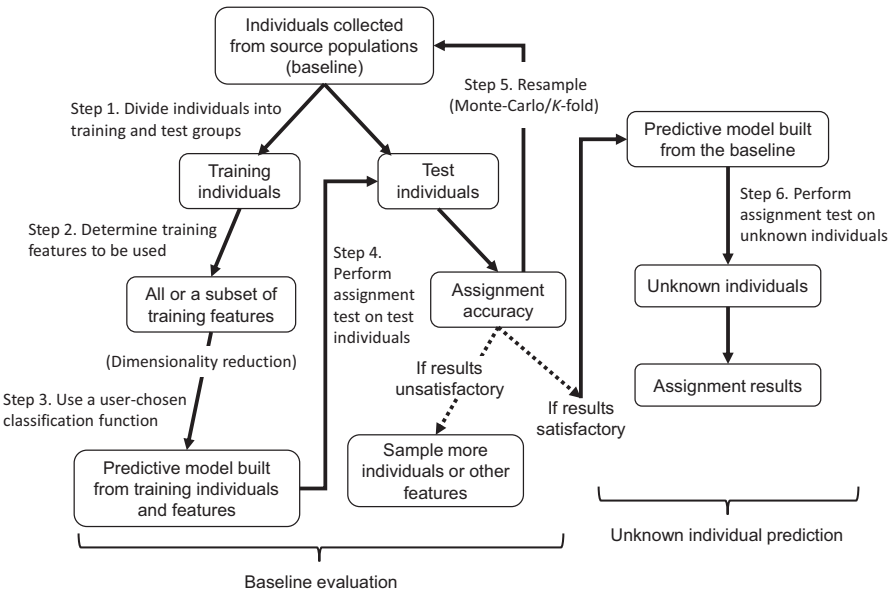

FIGURE 1 Analytical framework used in *assignPOP* in which baseline data were evaluated through resampling cross-validation. Various combinations of training individuals and features can be used to build predictive models and test on test individuals. When analysing genetic or genomic data, data dimensionality is reduced using principal component analysis. A baseline also can be used to estimate assignment accuracy and membership probability of individuals of unknown origins

However, the user can specify the number of PCs retained to fine tune and build predictive models in the function's argument.

2.5 | Classification models

The user can choose from five different classification machine-learning functions to build the predictive models (Step 3). These models are subsequently used to estimate membership probabilities of test individuals and assign individuals to a source population (determined by the greatest probability), both while evaluating the baseline data (Step 4) and conducting the assignment for unknown individuals (Step 6). The classification models include linear discriminant analysis from the package *MASS* (Ripley et al., 2016), support vector machines and naïve Bayes from the package *e1071* (Meyer et al., 2015), decision tree from package *tree* (Ripley, 2016) and random forest from the package *randomForest* (Cutler & Wiener, 2015).

2.6 | Data visualization

After the resampling cross-validation is completed, the user can calculate assignment accuracies with the function *accuracy.MC* and *accuracy.kfold* for results generated from *assign.MC* and *assign.kfold*, respectively. The predicted population of a test individual is determined by the highest membership probability, and if the predicted and original populations are identical, it is considered a correctly assigned individual. The assignment accuracies across the tests can be visualized in a box plot using the function *accuracy.plot*, whereas the results of membership probability can be visualized in a stacked-bar plot using the function *membership.plot*.

While the functions *assign.MC* and *assign.kfold* are used to conduct assignment during the baseline data evaluation phase (Step 1–5), the function *assign.X* is used to perform a one-time, final assignment test for the unknown individuals from a mixed population (Step 6), assuming the user is satisfied with the reliability of the baseline data for classification and assignment. The function *assign.X* provides the predicted source population of each individual and its posterior probability of membership to that population, as well as the others. All of these data are saved in text files, which can then be subsequently analysed, if so desired.

2.7 | Informative loci calling

Researchers may want to learn which genetic markers are most informative in the assignment tests, as this may allow them to design a set of targeted primers for fewer but more informative loci, thus saving time and money in genotyping. Our *assignPOP* offers this capability through the function *check.loci*. This function counts the occurrence of loci used in each of the Monte-Carlo and/or *K*-fold training datasets, and hence is useful when resampling subsets of high-F_{ST} loci as training loci. Because the F_{ST} value of a locus will likely vary when different individuals are sampled as training individuals, a locus that shows consistent high-F_{ST} values across the suite of training datasets is likely to be considered highly informative. The results of informative loci are shown in a table in which loci are ranked by the F_{ST} value in columns, and sorted by frequency in rows (an example can be found in our package tutorial website).

3 | EXAMPLES

Below, we provide hypothetical examples to demonstrate how baseline data are evaluated using a genetic dataset (1,000 SNP loci simulated by SIMCOAL2.0 (Laval & Excoffier, 2004)) and how assignment results can be improved by integrating the genetic and non-genetic datasets (four

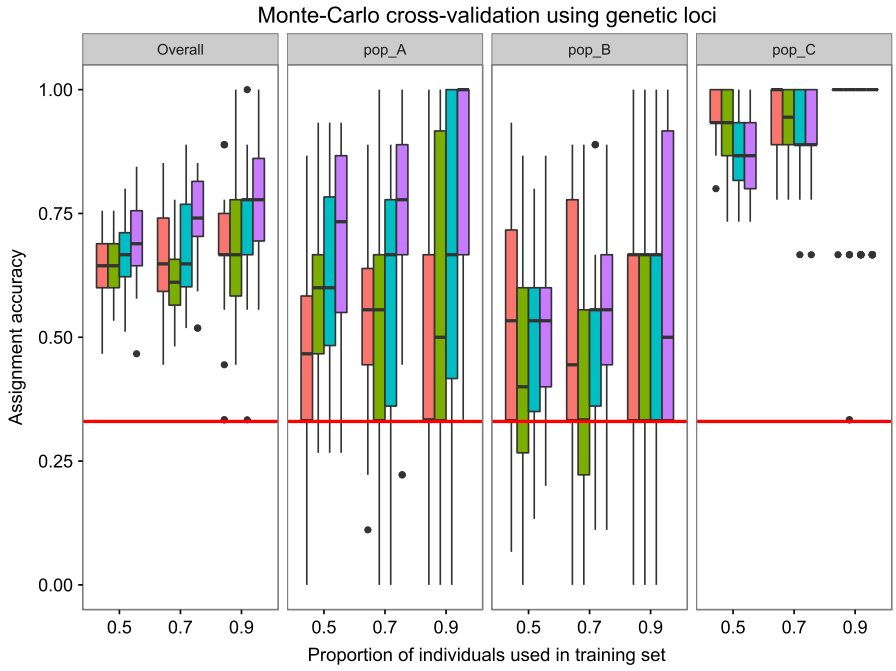

FIGURE 2 Assignment accuracies estimated via Monte–Carlo cross-validation and support vector machine methods, with sampling of three subsets of high-F_{ST} loci (top 10%, orange; top 25%, green; top 50%, blue) plus all loci (purple) crossed by three levels of training individuals, with 30 iterations. Box plot details: the line within the box is the median; top and bottom edges of the box are 25th and 75th percentiles; the ends of whiskers are the minimum and maximum of non-outliers, and outliers are shown as black circles. Horizontal red lines indicate the null population assignment rate, which was 33%

morphometric measurements) for three populations (A, B and C) of 30 individuals. In this simulated dataset, populations A and B have less genetic variation but have some degree of morphometric variation, whereas population C is both genetically and phenotypically different from the populations A and B. As such, we expect that genetic markers alone can be used to discriminate between groups A/B and C, but not between A and B, whereas the integrated markers can be used to discriminate among the three populations. The analytical workflow is as follows.

I. Import the library and genetic data

```
library(assignPOP)
YourGenepop <- read.Genepop( "simGenepop.txt", pop.names = c("pop_A","pop_B",
                             "pop_C"), haploid = FALSE)
#Download "simGenepop.txt" at https://goo.gl/ncDV2x
```

II. Remove low variance loci (optional). This step reduces the number of loci that have low variance across all individuals, and therefore could save computing time for further analyses, particularly when analysing large genomic data.

```
YourGenepopRd <- reduce.allele( YourGenepop, p = 0.95)
# p = 0.95 indicates the removal of loci having the frequency of an allele
greater than 0.95.
```

III. Perform Monte–Carlo cross-validation. The following script samples 50%, 70% and 90% of individuals from each population by the top 10%, 25%, 50% of high F_{ST} and all loci as training sets, and each training set combination is resampled 30 times. As a result, it performs a total of 360 assignment tests. In this example, support vector machine (model = "svm") classification model is used to build predictive models, and assignment results are saved under the folder named "Result-folder."

```
assign.MC( YourGenepopRd, dir="Result-folder/", train.inds=c(0.5,0.7,0.9),
           train.loci=c(0.1,0.25,0.5,1), loci.sample="fst", iterations=30,
           model="svm" )
```

IV. Calculate assignment accuracy of the Monte–Carlo cross-validation results. This step generates a dataset that includes assignment accuracies for each population across the tests, which can then be used to create assignment accuracy plots.

```
accuMC <- accuracy.MC( dir = "Result-folder/" )
```

V. Create an assignment accuracy box plot (Figure 2). The function *accuracy.plot* is used to create the plot. It is built with the *ggplot2* library (Wickham, 2016) so that the user can incorporate other *ggplot2* functions to modify the plot, as in the example script shown below.

```
library(ggplot2)
accuracy.plot( accuMC, pop=c("all", "pop_A", "pop_B", "pop_C")) +
  ylim(0, 1) + #Set y limit between 0 and 1
  annotate("segment",x=0.4,xend=3.6,y=0.33,yend=0.33,colour="red",size=1) +
#Add a red horizontal line at y = 0.33 (null assignment rate for 3 population
s)
  ggtitle("Monte-Carlo cross-validation using genetic loci")+
#Add a plot title
  theme(plot.title = element_text(size=20, face="bold"))
#Edit plot title text size
```

VI. Perform *K*-fold cross-validation. The following script divides individuals from each population into 3, 4 or 5 groups and randomly samples 10%, 25% or 50% of loci or uses all loci as training data, resulting in a total of 48 assignment tests. In this example, linear discriminant function analysis (model = "lda") is used to build

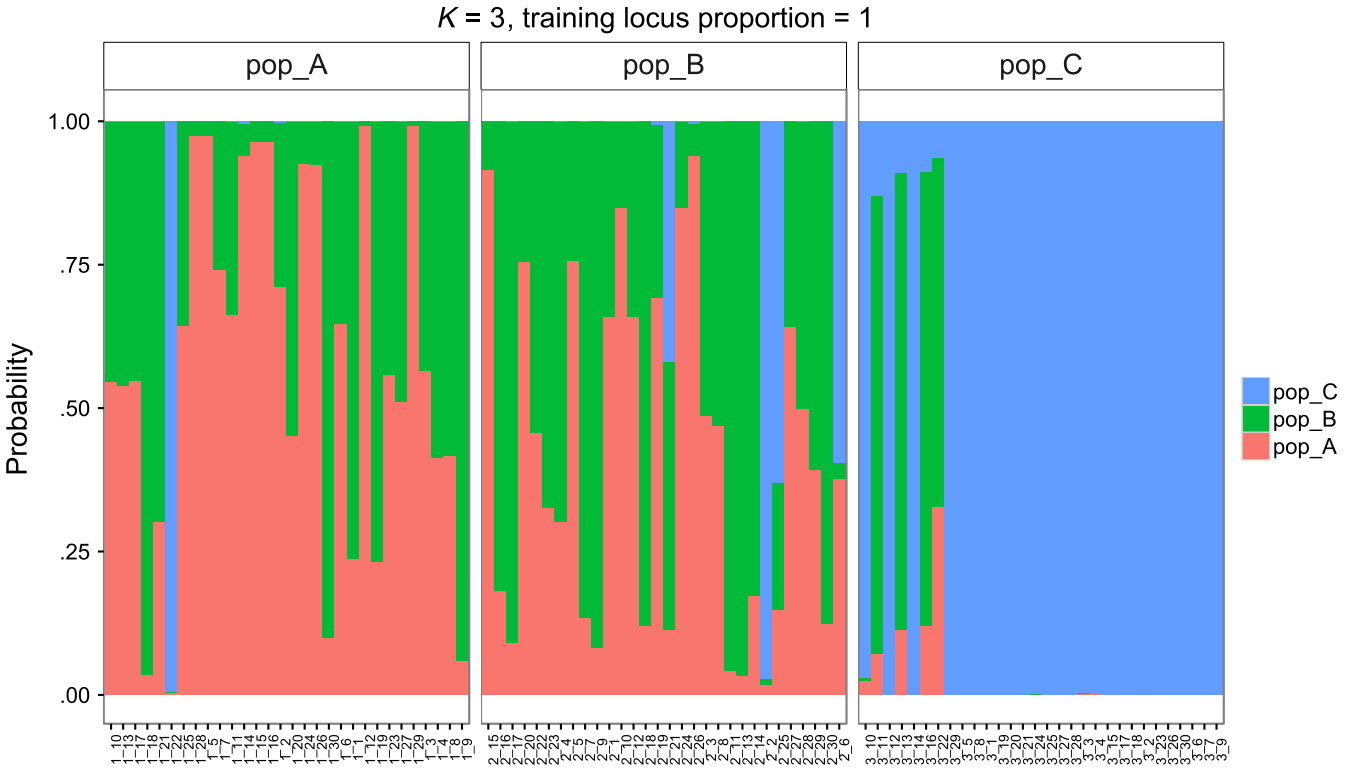

FIGURE 3 Membership probability of three hypothetical populations. Probabilities of individuals were estimated using 3-fold cross-validation and all loci (1,000 SNPs)

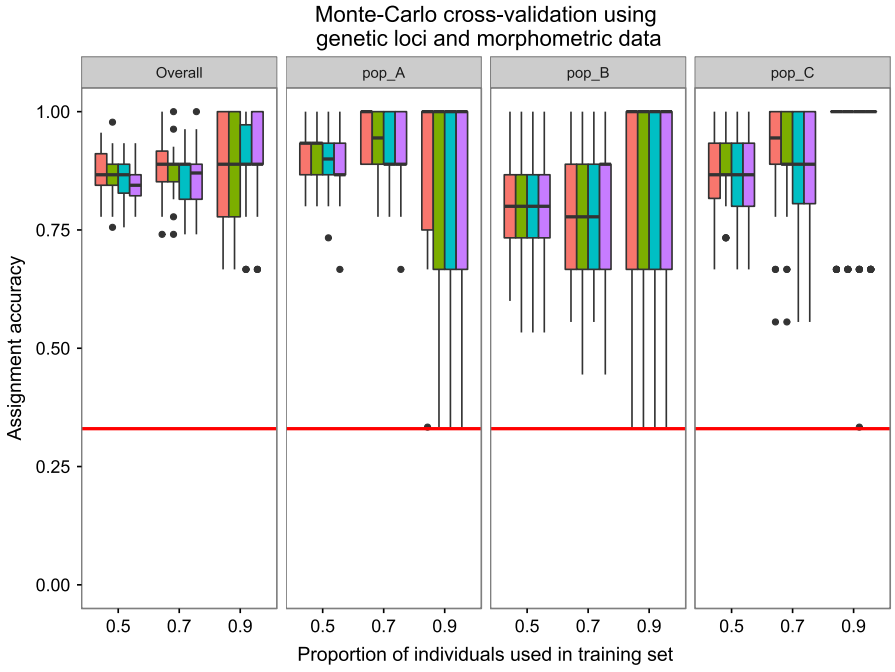

FIGURE 4 Assignment accuracies estimated via Monte–Carlo cross-validation and support vector machine methods, with random sampling of three levels of training individuals crossed by four levels of training loci (each box cluster from left to right: 10%, 25%, 50%, all) plus four morphometric measurements crossed by 30 iterations. See Figure 2 for box plot description details

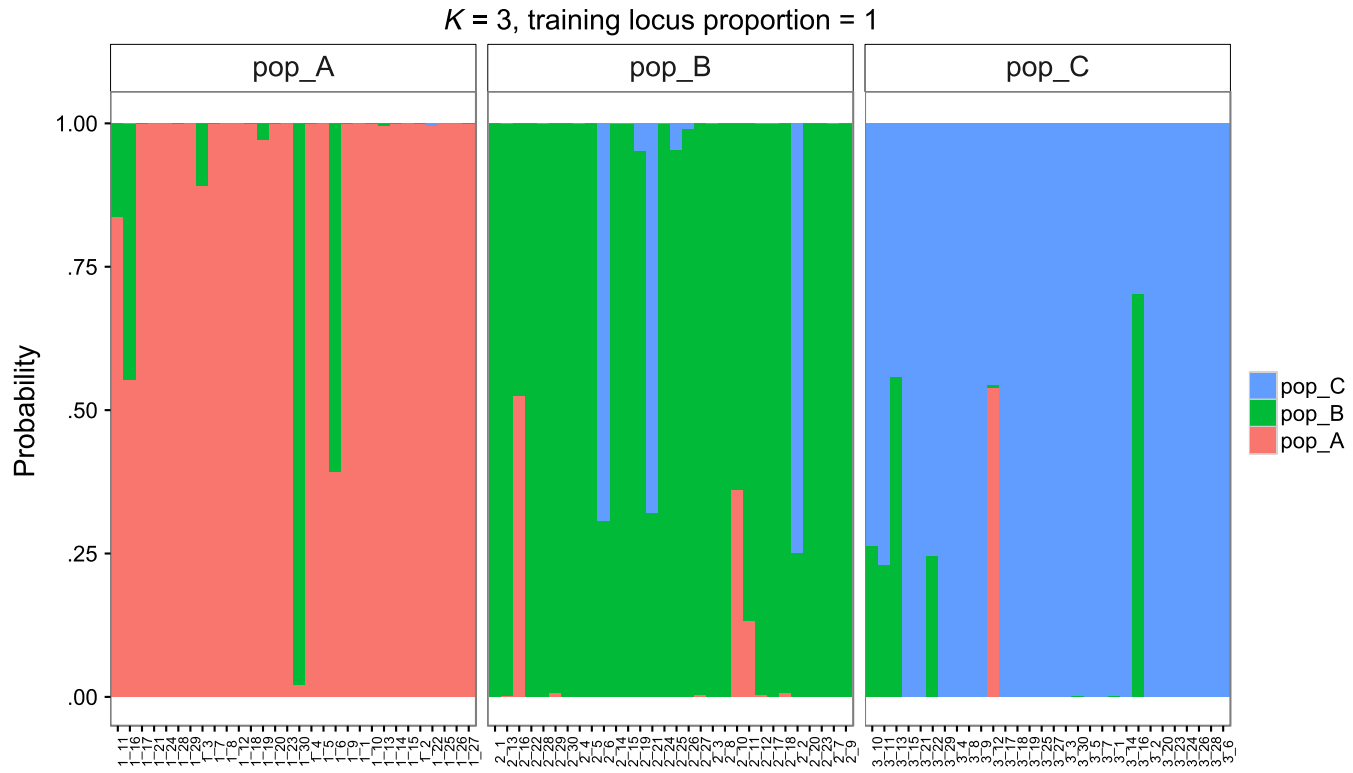

FIGURE 5 Membership probability of three hypothetical populations. Probabilities of individuals were estimated using 3-fold cross-validation and 1,000 SNPs plus four morphometric measurements

predictive models, and assignment results are saved under the folder named "Result-folder2." The results are used to create membership probability plots.

```
assign.kfold( YourGenepopRd, k.fold=c(3,4,5), train.loci=c(0.1,0.25,0.5,1),
              loci.sample="random", dir="Result-folder2/", model="lda" )
```

VII. Create a membership probability plot (Figure 3). After entering the following script, the user is prompted to specify which assignment results (*K* groups and training loci) and styles should be used to create the plot. Here, we created the plot using the results from *K* = 3 and all loci in the training data.

```
membership.plot( dir = "Result-folder2/" )
```

VIII. Concatenate genetic and morphometric data. The script below concatenates genetic (YourGenepopRd) and non-genetic (morphData.csv) data into a new integrated dataset that can be used to perform the baseline evaluation.

```
YourIntegrateData <- compile.data( YourGenepopRd, "morphData.csv" )
#Download "morphData.csv" at https://goo.gl/WTx7cA
```

The object (YourIntegrateData) can be further analysed using the same analytical workflow (see Steps III–VII) to create the plots of assignment accuracy (Figure 4) and membership probability (Figure 5). Compared with the results from using genetic data alone, the integrated dataset resulted in higher assignment accuracies in both populations A and B (Figure 4). Also, more individuals were correctly assigned to populations A and B (Figure 5). As expected, the overall assignment results were improved by integrating genetic and morphometric data. It is worth noting that mean assignment accuracy increased with increasing proportion of the data used in the training set, with the variance also increasing in some instances (see populations A and B in Figure 5). These results suggest that assignment accuracy could become upwardly biased when the proportion of individuals used as training data is too large (e.g., leave-one-out cross-validation approach). Lastly, after baseline data are evaluated and the results are deemed satisfactory, users can perform an assignment test on a mixture of individuals from unknown origins using the function *assign.X*. Example codes and datasets used here can be found on our tutorial website (http://alexkychen.github.io/assignPOP).

4 | CONCLUSIONS

Herein, we described the first release of *assignPOP*, which includes several novel features to perform population assignment using the concept and methods of machine learning. It is a versatile R package in that it allows both genetic and non-genetic data to be integrated. In turn, researchers can compare assignment results generated from the different data types, thus enabling them to more easily identify useful markers for population discrimination than existing tools. We show that this determination is important because, if population-specific features exist in these different data types, their integration can improve overall population assignment results. Moreover, *assignPOP*'s ability to resample various combinations of training data and perform independent cross-validation tests can help researchers evaluate whether a baseline dataset has sufficient discriminatory power to predict the source population(s) of individuals while simultaneously preventing the predictive models from overfitting. By facilitating the ability of users to explore population structure, identify markers that can best discriminate among populations, and then use these population-specific markers to determine the source origin(s) of individuals from a "mixed" population, *assignPOP* is likely to benefit both species conservation and management.

ACKNOWLEDGEMENTS

We thank Dr. Laura Kubatko for providing suggestions to data analyses and reviewing the earlier version of the manuscript. This study was supported by the Federal Aid in Sport Fish Restoration Program (F-69-P, Fish Management in Ohio), administered jointly by the US Fish and Wildlife Service and the Ohio Division of Wildlife (FADR68 to S.A.L., E.A.M. and K.-Y.C.). We have no conflict of interest to declare.

AUTHORS' CONTRIBUTIONS

S.A.L., H.L.G., and E.A.M. conceived the ideas and proposed methodology; K.-Y.C., M.G.S., and A.C.F. designed the methodology; K.-Y.C. developed the software and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

R code and data used in this manuscript are available through the R package *assignPOP*, which can be downloaded via CRAN (https://CRAN.R-project.org/package=assignPOP) or Github (https://github.com/alexkychen/assignPOP). A demo script is uploaded as a supplementary file in Supporting Information.

ORCID

Kuan-Yu Chen http://orcid.org/0000-0001-9904-0886
Elizabeth A. Marschall http://orcid.org/0000-0002-8026-4203
H. Lisle Gibbs http://orcid.org/0000-0001-7461-3393
Stuart A. Ludsin http://orcid.org/0000-0002-3866-2216

REFERENCES

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: Standard methods are upwardly biased. *Molecular Ecology Resources*, *10*, 701–710.

Becker, B. J., Levin, L. A., Fodrie, F. J., & McMillan, P. A. (2007). Complex larval connectivity patterns among marine invertebrate populations. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 3267–3272.

Begg, G. A., & Waldman, J. R. (1999). An holistic approach to fish stock identification. *Fisheries Research*, *43*, 35–44.

Bradbury, I. R., Campana, S. E., & Bentzen, P. (2008). Estimating contemporary early life-history dispersal in an estuarine fish: Integrating molecular and otolith elemental approaches. *Molecular Ecology*, *17*, 1438–1450.

Cadrin, S. X. (2000). Advances in morphometric identification of fishery stocks. *Reviews in Fish Biology and Fisheries*, *10*, 91–112.

Cadrin, S. X., Kerr, L. A., & Mariani, S. (2013). *Stock identification methods: Applications in fishery science*. London, UK: Elsevier Academic Press.

Campana, S. E., Fowler, A. J., & Jones, C. M. (1994). Otolith elemental fingerprinting for stock identification of Atlantic cod (*Gadus morhua*) using laser ablation ICPMS. *Canadian Journal of Fisheries and Aquatic Sciences*, *51*, 1942–1950.

Chabot, A. A., Hobson, K. A., Wilgenburg, S. L. V., McQuat, G. J., & Lougheed, S. C. (2012). Advances in linking wintering migrant birds to

their breeding-ground origins using combined analyses of genetic and stable isotope markers. *PLoS ONE*, *7*, e43627.
Clegg, S. M., Kelly, J. F., Kimura, M., & Smith, T. B. (2003). Combining genetic markers and stable isotopes to reveal population connectivity and migration patterns in a Neotropical migrant, Wilson's warbler (*Wilsonia pusilla*). *Molecular Ecology*, *12*, 819–830.
Cutler, F., & Wiener, R. (2015). *RandomForest: Breiman and Cutler's random forests for classification and regression*.
Czesny, S., Dabrowski, K., Christensen, J. E., Van Eenennaam, J., & Doroshov, S. (2000). Discrimination of wild and domestic origin of sturgeon ova based on lipids and fatty acid analysis. *Aquaculture*, *189*, 145–153.
Gibbs, H. L., Weatherhead, P. J., Boag, P. T., White, B. N., Tabak, L. M., & Hoysak, D. J. (1990). Realized reproductive success of polygynous red-winged blackbirds revealed by DNA markers. *Science*, *250*, 1394.
Gómez-Díaz, E., & González-Solís, J. (2007). Geographic assignment of seabirds to their origin: Combining morphologic, genetic, and biogeochemical analyses. *Ecological Applications*, *17*, 1484–1498.
Guillot, G., Renaud, S., Ledevin, R., Michaux, J., & Claude, J. (2012). A unifying model for the analysis of phenotypic, genetic, and geographic data. *Systematic Biology*, *61*, 897–911.
Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*, 149–161.
Hardesty, B. D., Hubbell, S. P., & Bermingham, E. (2006). Genetic evidence of frequent long-distance recruitment in a vertebrate-dispersed tree. *Ecology Letters*, *9*, 516–525.
Hellberg, M. E., Burton, R. S., Neigel, J. E., & Palumbi, S. R. (2002). Genetic assessment of connectivity among marine populations. *Bulletin of Marine Science*, *70*, 273–290.
Hobson, K. A. (1999). Tracing origins and migration of wildlife using stable isotopes: A review. *Oecologia*, *120*, 314–326.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Resampling methods. An introduction to statistical learning. In *Springer texts in statistics* (pp. 175–201). New York, NY: Springer.
Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*, 1403–1405.
Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*, 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*, 94.
Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.
Kelly, J. F., Ruegg, K. C., & Smith, T. B. (2005). Combining isotopic and genetic markers to identify breeding origins of migrant birds. *Ecological Applications*, *15*, 1487–1494.
Koehler, A. V., Pearce, J. M., Flint, P. L., Franson, J. C., & Ip, H. S. (2008). Genetic evidence of intercontinental movement of avian influenza in a migratory bird: The northern pintail (*Anas acuta*). *Molecular Ecology*, *17*, 4754–4762.
Laval, G., & Excoffier, L. (2004). SIMCOAL 2.0: A program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, *20*, 2485–2487.
Mackenzie, K. (2002). Parasites as biological tags in population studies of marine organisms: An update. *Parasitology*, *124*, 153–163.
Manel, S., Gaggiotti, O. E., & Waples, R. S. (2005). Assignment methods: Matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, *20*, 136–142.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*. TU Wien.
Perrier, C., Daverat, F., Evanno, G., Pécheyran, C., Bagliniere, J.-L., & Roussel, J.-M. (2011). Coupling genetic and otolith trace element analyses to identify river-born fish with hatchery pedigrees in stocked Atlantic salmon (*Salmo salar*) populations. *Canadian Journal of Fisheries and Aquatic Sciences*, *68*, 977–987.
Piry, S., Alapetite, A., Cornuet, J.-M., Paetkau, D., Baudouin, L., & Estoup, A. (2004). GENECLASS2: A software for genetic assignment and first-generation migrant detection. *Journal of Heredity*, *95*, 536–539.
Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.
Puechmaille, S. J. (2016). The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*, 608–627.
Ripley, B. (2016). *tree: Classification and regression trees*.
Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2016). *MASS: Support functions and datasets for Venables and Ripley's MASS*.
Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity analysis of K-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 569–575.
Ross, K. G. (2001). Molecular ecology of social behaviour: Analyses of breeding systems and genetic structure. *Molecular Ecology*, *10*, 265–284.
Rousset, F. (2008). Genepop'007: A complete re-implementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*, 103–106.
Ruegg, K. C., Anderson, E., Harrigan, R. J., Paxton, K. L., Kelly, J., Moore, F., & Smith, T. B. (2016). Identifying migrant origins using genetics, isotopes, and habitat suitability. *bioRxiv*, 085456.
Rundel, C. W., Wunder, M. B., Alvarado, A. H., Ruegg, K. C., Harrigan, R., Schuh, A., … Novembre, J. (2013). Novel statistical methods for integrating genetic and stable isotope data to infer individual-level migratory connectivity. *Molecular Ecology*, *22*, 4163–4176.
Smith, S. J., & Campana, S. E. (2010). Integrated stock mixture analysis for continuous and categorical data, with application to genetic–otolith combinations. *Canadian Journal of Fisheries and Aquatic Sciences*, *67*, 1533–1548.
Wang, J. (2016). The computer program Structure for assigning individuals to populations: Easy to use but easier to misuse. *Molecular Ecology Resources*, https://doi.org/10.1111/1755-0998.12650
Waples, R. S. (2010). High-grading bias: Subtle problems with assessing power of selected subsets of loci for population assignment. *Molecular Ecology*, *19*, 2599–2601.
Waples, R. S., & Gaggiotti, O. (2006). Invited review: What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, *15*, 1419–1439.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.
Xu, Q.-S., & Liang, Y.-Z. (2001). Monte–Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*, 1–11.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Chen K-Y, Marschall EA, Sovic MG, Fries AC, Gibbs HL, Ludsin SA. assignPOP: An R package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework. *Methods Ecol Evol*. 2018;9:439–446. https://doi.org/10.1111/2041-210X.12897