BIG DATA ANALYTICS
in Molecular Ecology

MSc Bioinformatics · Module 5 · Feb 2023

https://mscbioinformatics.uab.cat

UAB
Universitat
Autònoma
de Barcelona

Relevance of Ecotype Components in the Genetic Population Structure of Killer Whales

Alberto Carrasco[1]

[1] MSc in Bioinformatics, Autonomous University of Barcelona

Abstract

Killer whale (*Orcinus orca*) populations represent a good model for studying the relationship between ecological behaviors and genetic diversity. Despite their substantial dispersal abilities, these subpopulations have evolved into different ecotypes due to exploiting narrow ecological niches. Focusing on the North Pacific and other ecotypes, several clustering methods and a random forest model are applied to a dataset of genetic variants in order to ascertain the importance of ecotype on the genomic divergence of these populations. The findings suggest that cultural transmission plays an essential role in the population structure of killer whales, even when gene flow and mobility are not restricted.

Keywords: evolution, clustering, pca, machine learning, population genomics, ecological genetics

Introduction

Unraveling the patterns of population structure in animal species is key in order to expand our knowledge on how species adapt and diverge. In fact, studying which factors contribute the most to the underlying genetic structure of natural populations is critical for making an accurate evaluation of their status and thus aiding conservation efforts [1]. Some of these factors, such as dispersal and gene flow, have been studied for decades and their impact on population structure is more or less detailed. However, to which extent the culture of animals can influence the evolution of their genomes is yet to be fully understood. Filling this gap is important since the transmission of specific behaviors can serve as a strategy to adapt to new environments and also create an opportunity for natural selection to act on adaptive genomic variants [2].

The interplay between ecological, cultural, and genetic variation has the potential to create diversity within species, and killer whales (*Orcinus orca*) provide a great framework to test several hypotheses on this matter [2]. Despite their substantial dispersal abilities as a marine species, killer whale subpopulations have evolved due to exploiting narrow ecological niches over small geographic areas [3]. This has led to the appearance of several ecotypes with significant genetic differences. These ecological categories or ecotypes are defined mainly on prey preference and are preserved through social education from mothers to their offspring. Nonetheless, several other features can tell ecotypes apart such as phenotype, mating system, acoustic patterns, and social dynamics. Furthermore, this differentiation can not only be observed between groups which are isolated by distance but also between sympatric ecotypes [2]. As a matter of fact, killer whale populations inhabiting the North Pacific serve as an interesting model to study the relevance of cultural aspects on dispersal since three different ecotypes ("resident", "transient", "offshores") have been described to live there in sympatry [3].

Those populations described as "resident" feed mainly on fish species while "transient" killer whales rather prey on marine mammals [4]. The food source of "offshores" populations was a mystery for decades, but recent studies suggest a fish diet focused on shark species. Other instances of ecotypes are the "Antarctic type B" which feed on penguins, fish and seals, and the populations living around Iceland known to prey on herring although without a defined feeding behavior yet.

Here, the relevance of ecological behaviors versus geographical proximity on the population structure of killer whales despite no prior restrictions on gene flow and mobility is tested. To do so, several clustering methods are applied to the principal components which explain the variance observed in a big dataset of single-nucleotide polymorphic (SNP) variants corresponding to different populations (N=9) and ecotypes (N=5). In these samples, some sympatric populations show differing ecotypes and some of the ecotypes are shared by distant populations. Moreover, a random forest model is trained in order to improve the accuracy of assigning samples to a certain population without knowing their ecotype.

Materials and Methods

Sample data

Samples correspond to a DNA archive with submissions from several previous studies. All North Pacific samples (AT, AR, SR, CT, OS, BS, RU) were collected by Hoelzel *et al.* in 2007 [4] as well as the Iceland samples (IC) that act as an outgroup. Additional samples from the Antarctic region of Marion Island in South Africa (MI) are also included as an outgroup, retrieved from Moura *et al.* [3]. In summary, a total of 115 individuals belonging to different populations and ecotypes are studied. Details on sample size, ecotype and areas are described in **Table 1**. Samples from Alaska (AR, AT) and the American West Coast (SR, CT) represent instances of different ecotypes (resident, transient) occurring in sympatry (**Figure 1**).

For the analysis, restriction-site-associated DNA (RAD) single-nucleotide polymorphic (SNP) markers are the only variants used. All the SNPs included (N=3678) are present and have scored accurately for all individuals (see **Data Availability**).

Unsupervised methods

Unsupervised learning methods are used ignoring the labels of our data. This is useful to gain insight into the structure of the populations and identify patterns that may not be immediately apparent.

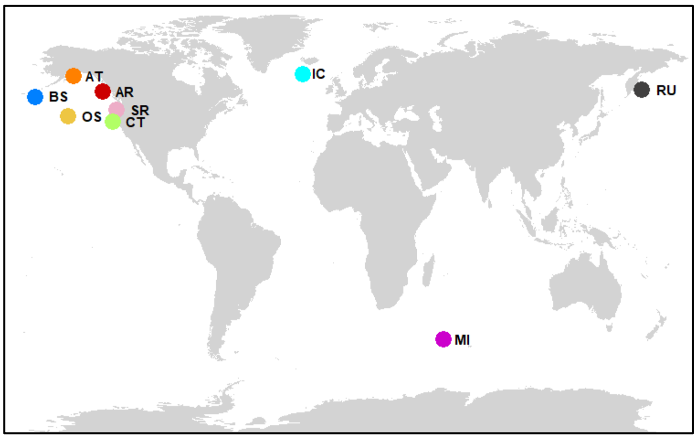

Figure 1. World map showing sample locations by population code.

1. Dimensionality reduction

The number of features in the dataset (3678) is too high for analysis and visualization. Therefore, the dimensions need to be restricted. This is done by transforming the original features into a new set with the most important information by projecting the data onto a lower-dimensional space. Dimensionality reduction decreases the complexity and helps to improve the performance and visualization of machine learning methods.

Here, two dimensionality reduction methods are performed in R and then compared. Principal Component Analysis (PCA) is a linear technique which involves calculating the covariance matrix of the data and then computing its eigenvectors and eigenvalues. After identifying the directions of maximum variance in the data, it is projected onto these directions. The t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear technique which minimizes the divergence between a high-dimensional distribution of pairwise similarities between data points, and a low-dimensional distribution of similarities between the corresponding points in the low-dimensional space.

Table 1. Number of samples per population, corresponding ecotype, and sampling location.

Population Code	Ecotype	Sampling Location	Number of samples
AR	Resident	Alaska, USA	17
SR	Resident	Washington State, USA	13
RU	Resident	Kamchatka, Russia	9
BS	Resident	Bering Sea	13
AT	Transient	Alaska, USA	21
CT	Transient	California, USA	16
OS	Offshore	Eastern North Pacific	7
IC	Undetermined	Iceland	6
MI	Antarctic Type B	Marion Island, SA	13

2. Clustering methods

Clustering is a technique used to group data points with similar characteristics in order to pinpoint clusters within the data. The input for these methods is the output data of PCA with reduced dimensionality in order to increase the yield and visualization of the clustering results.

In this analysis, two clustering algorithms are applied to the principal components of the dataset in R (`cluster` package). Both algorithms are used using two different numbers of clusters, k=5 and k=9. These numbers represent the number of ecotypes and populations in our samples, respectively. The aim is to compare both algorithms in order to check which performs best with our data. Also, a comparison of both clustering parameters tests whether these algorithms are able to assign individuals to a cluster which accurately represents their ecotype/population. To test the precision of each algorithm, the confusion matrices for each method and number of k are created. Moreover, the Adjusted Rand Index (ARI) is computed to measure the similarity between clustering results by using the `mclust` package in R.

K-means clustering works by randomly selecting k data points as centroids and then iterating assigning every data point to the closest centroid while updating the centroids based on the mean of the data points assigned to each cluster. The goal is to minimize the sum of squared distances between data points and their assigned centroid. Here, it is applied using 100 random initializations as a hyperparameter to increase the chance of finding a good solution. Hierarchical clustering creates a hierarchy of clusters by recursively partitioning the data into smaller subclusters. As a results, a dendrogram illustrating hierarchical relationships between clusters is generated. This method is run on the Euclidean distance matrix of the PCA data.

Supervised methods

Supervised learning consists in training an algorithm on a set of labeled input-output pairs in order to learn a function that can predict accurately the output for new input data never seen before.

1. Random forest

Random Forest can be used for classification tasks, it is an ensemble learning method which combines multiple decision trees to improve its predictive performance. Each tree is trained on a random subset of the training data and a random subset of the input features. This helps to reduce overfitting and improve generalization by introducing randomness into the model. Finally, the predictions of all trees are combined to generate a final prediction.

In this study, a random forest model is trained for assigning the correct population to samples. In order to do so, the package `randomForest` in R is used. First, data is randomly split in two portions, 80% corresponding to the training dataset and 20% corresponding to the testing dataset. Then, the best mtry for the model is estimated using the `tuneRF()` function in R. This is an important hyperparameter that specifies the maximum number of features considered at each split and controls the randomness and generalization performance of the model. In fact, adjusting mtry balances the model and improves its predictive performance. Next, the random forest is trained to identify killer whale populations using 20,000 trees, and this training is used to predict the feature in the testing data. Finally, the accuracy of the model and the prediction is measured by a confusion matrix. The area under the ROC curve (AOC) is also estimated for summarizing the model's performance.

Results

Dimensionality reduction

Principal Components Analysis (PCA) successfully reduces the features of the dataset (SNP genotypes) to only 115 principal components. Out of these, PC1 and PC2 explain the most variance in the data, 17.7 and 6.6% respectively. The t-SNE algorithm is also able to reduce the dimensionality of the dataset after using a perplexity of 38 and a theta of 0. Both hyperparameters have a significant impact on the quality of the result and their optimal values for this specific dataset have been chosen using trial and error. The visualization of the two outputs allows for a comparison of the performance of both methods (**Figure 2**). The result of the two algorithms shows populations can be discriminated as a function of their ecotype instead of their population or geographic proximity. The results are reliable as the outgroups (IC, MI) and the three Northern Pacific ecotypes (resident, transient, offshores) are separated accurately. Despite PCA and t-SNE work well with our data, PCA reduction generates a sharper distinction of the samples and therefore it is selected as input for the subsequent clustering methods.

Clustering methods

Both K-means and Hierarchical clustering algorithms are applied to the PCA data with reduced dimensionality which results in a better computing performance. Two different numbers of predefined cluster are used, k=5 and k=9, in order to test how good clustering methods can tell apart the samples based on either ecotype or population. As seen in **Figure 3**, both methods show a similar performance in either predefined k value. In fact, the adjusted Rand index (ARI) for the k=5 outputs is 1, and the ARI for the k=9 tests is 0.89.

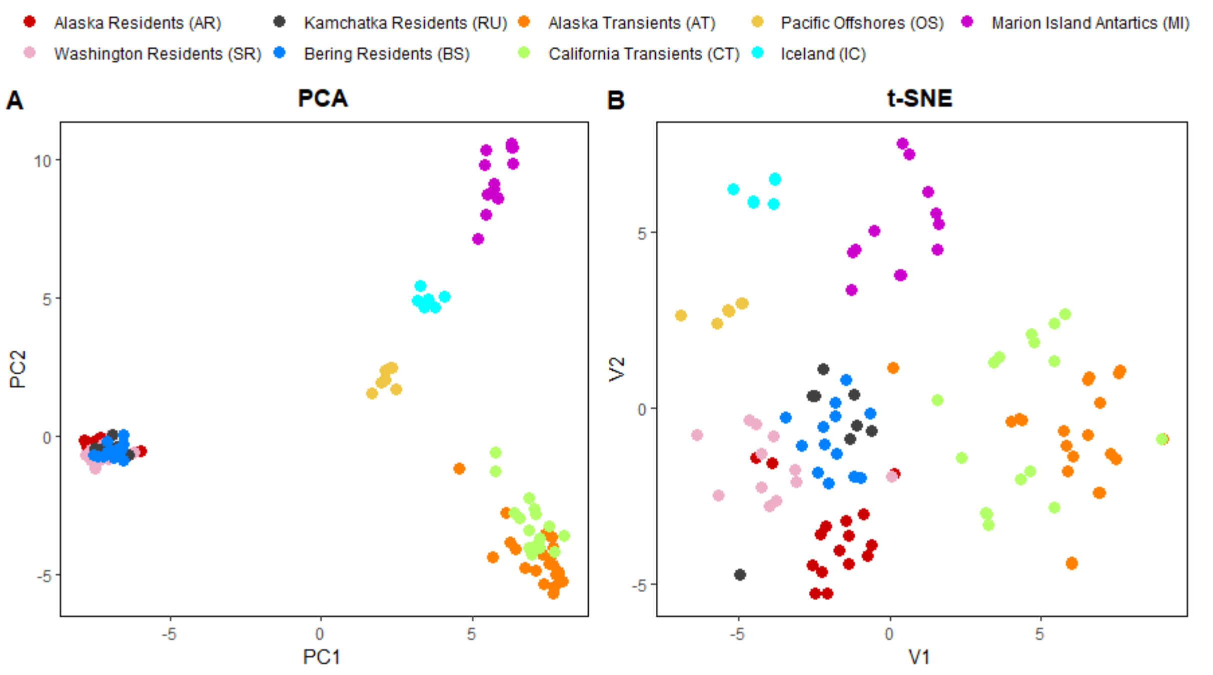

Figure 2. Visualization of the results of dimensionality reduction, using PCA (2A) and t-SNE (2B). Samples are colored according to their population of origin.

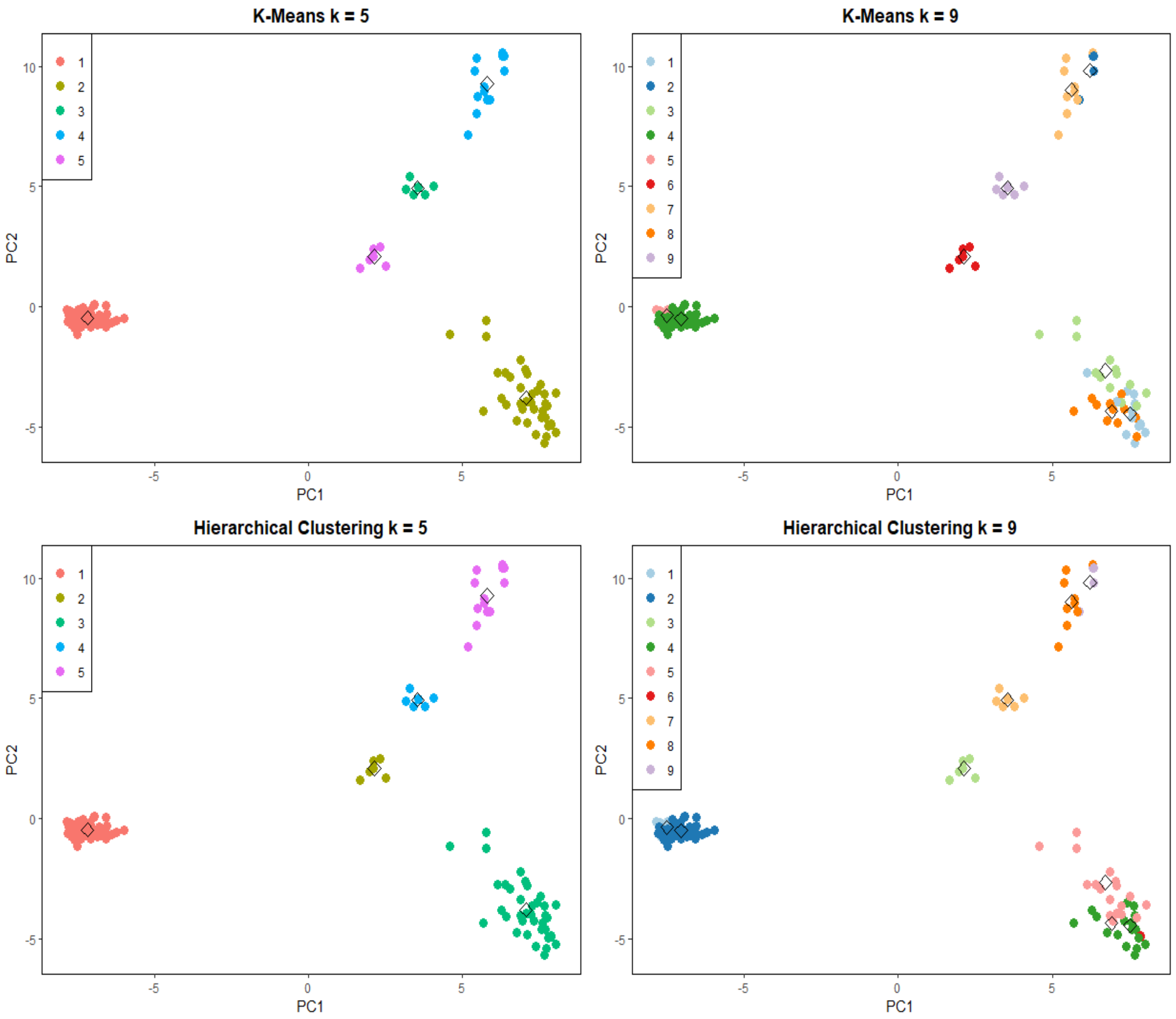

Figure 3. Visualization of the clustering methods results for *k*=5 and *k*=9. Top row corresponds to K-means clustering outputs and bottom row indicates the outputs for the Hierarchical clustering.

Both methods are able to correctly cluster all ecotypes but fail to fully identify the nine different populations in the dataset. For *k*=5, the cluster-ecotype distribution is perfect for the two algorithms, while *k*=9 shows a lower accuracy (confusion matrices are available in the **Supplementary Data Tables S1-3**) but performs slightly better in Hierarchical clustering when identifying the transient populations (AT, CT). In the *k*=9 clusterings, the four resident populations are lumped together in only two clusters while the two transient subsets are split in three. Interestingly, the Antarctic outgroup is identified as two different putative populations that may correspond to the Antarctic B1 and B2 types [2] or just be an artifact. Both Icelandic and offshore populations/ecotypes are clustered adequately in the four tests.

Random forest

Given the poor performance of clustering methods for correctly identifying subpopulations instead of ecotypes, a random forest model is created and trained to improve the yield of this task. After splitting the PCA dataset in training (80%) and testing (20%) subsets, the random forest model is trained using 20,000 trees to increase the accuracy. In our case, a value of mtry=9 is estimated (see **Figure S1**). The trained random forest model has an OOB estimate of error rate of 17.39% (confusion matrix available in the **Supplementary Data Table S4**). After training, this model is applied to the testing dataset to check its performance. As a result, the model proves itself as a better tool to predict populations than clustering methods since it is able to achieve an accuracy of 91.3% and AUC of 0.94. It still struggles slightly telling apart three of the four resident populations from the North Pacific (AR, BS, and SR) (**Table 2**).

Table 2. Statistics by population for the trained random forest model (sensitivity, specificity, positive predictive value, and negative predictive value).

Stat.	AR	AT	BS	CT	IC	MI	OS	RU	SR
Sens.	0.50	1	0.67	1	1	1	1	1	1
Spec.	1	1	1	1	1	1	1	1	0.91
Pos.	1	1	1	1	1	1	1	1	0.33
Neg.	0.95	1	0.95	1	1	1	1	1	1

Discussion

The effects of factors such as dispersal and consanguinity on population structure have been thoroughly discussed in ecological research, but the impact of cultural behaviors in the genotype distribution of species has yet to be fully understood. Here, the genetic structure of several killer whale subpopulations is analyzed to address this question. These populations have no *a priori* mobility and gene flow restrictions but are divided in ecological groups, ecotypes which have been described to have significant genetic differences. Therefore, several machine learning methods are applied to discern the contribution of ecological factors to the distribution of SNP genotypes versus the impact of geographical proximity.

The dimensionality reduction of the data shows a genetic structure heavily based on ecotypes. The principal components of the reduced dataset distribute the nine populations according to their ecotype and not their — geographic proximity (Alaskan residents are not grouped along Alaskan transients). Both clustering methods, K-means and Hierarchical clustering, also create an accurate grouping of the ecotypes present in the dataset. In fact, this can be seen not only in the split of Alaskan killer whales (AR, AT) but also in a more global scale in the separation of the Eastern North Pacific killer whales (OS) from the rest of North Pacific populations (AR, SR, RU, BS, AT, CT) despite inhabiting a common area due to their distinct ecotype (offshore). The clustering methods did not perform well assigning the samples to unique populations. The four resident populations are classified in just two groups, the two transients in three, and the Antarctic outgroup is split in two (this result might be revealing the identification of type B1 and type B2 individuals within the same Antarctic type B sample but this needs further evaluation to be concluded). Altogether, the clustering results reveal a prevalence of the contribution of ecological and behavioral factors over the geographical proximity in the population structure of killer whales.

Moreover, a random forest is trained to improve the performance of assigning the samples to one of the nine populations. Using 20,000 trees, the model is able to reach a high accuracy of 91.3% and assigns most individuals to their correct population. This reveals that ecotypes influence the genetic makeup of killer whale populations, but they still retain some unique features characteristic of each group. Yet, it struggles telling apart three of the four resident populations (AR, BS, SR) which hints to a closer interbreeding among these.

In summary, the results reveal that ecotypes are a deciding factor in the genetic population structure of killer whales while geographical proximity is not significant. Those populations living in sympatry seem not to be genetically similar if they belong to different ecotypes (for instance, resident and transient killer whales in Alaska) despite there being no barriers for dispersal. This indicates that foraging strategies do influence social behavior and this, in turn, affects population structure by restricting gene flow. Populations with different ecotypes seem to not interbreed between each other. It is important to highlight the possible contribution of other evolution factors to this differentiation among ecotypes beyond the lack of reproduction. Since the prey sources are diverse, they require significantly different strategies to be successfully obtained. Hence, natural selection acts in a similar way on the genetic variation of the subpopulations belonging to the same ecotype and thus these will most likely share the same genetic variants. Genetic drift is also a probable factor contributing to the differentiation between populations within and without the same ecotype. In future studies, a bigger sample dataset including more types of variants (indels, CNV) is required to elucidate a more accurate contribution of ecological factors and evolution forces to the population structure of killer whales.

Data Availability

The raw DNA sequences for all samples in this study are deposited at GenBank with accessions: SAMN03020306–SAMN03020378; SAMN02820869–SAMN02820892; SAMN02820894–SAMN02820911. The VCF file containing SNP genotype information for all samples can be found at this repository.

The source code in R used to perform the analyses described in this paper (https://github.com/albcp2).

Acknowledgements

I want to thank all the researchers from the School of Biological and Biomedical Sciences of Durham University and the Institute of Integrative Biology in Liverpool (UK), the Department of Zoology and Entomology of the University of Pretoria (SA) and the National Marine Mammal Laboratory in Seattle (USA) for carrying out and funding all the projects which yielded the genomic data utilized in this paper and making it public access for everyone to use.

REFERENCES

[1] Parsons, Kim E *et al.* Geographic Patterns of Genetic Differentiation among Killer Whales in the Northern North Pacific. *Journal of Heredity* 104, 737 (2013). https://doi.org/10.1093/jhered/est037

[2] Foote, A *et al.* Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat Commun* 7, 11693 (2016). https://doi.org/10.1038/ncomms11693

[3] Moura, Andre E *et al.* Population genomics of the killer whale indicates ecotype evolution in sympatry involving both selection and drift. *Molecular Ecology* 23, 5179 (2014). https://doi.org/10.1111/mec.12929

[4] Hoelzel, A R *et al.* Evolution of Population Structure in a Highly Social Top Predator, the Killer Whale. *Molecular Biology and Evolution* 24, 1407 (2007). https://doi.org/10.1093/molbev/msm063

Supplementary Data

Table S1. Confusion matrix for *k*=5 for both K-means and Hierarchical clustering methods showing precision values for all ecotypes.

Ecotype	1	3	4	5	2	Total	Precision
Resident	52	0	0	0	0	52	100%
Transient	0	37	0	0	0	37	100%
Undetermined	0	0	6	0	0	6	100%
Antarctic Type B	0	0	0	13	0	13	100%
Offshore	0	0	0	0	7	7	100%

Table S2. Confusion matrix for *k*=9 for K-means showing precision values for all populations.

Population	5	1	4	3	9	7	6	8	2	Total	Precision
AR	13	0	4	0	0	0	0	0	0	17	76.47%
AT	0	12	0	1	0	0	0	8	0	21	57.14%
BS	0	0	13	0	0	0	0	0	0	13	100%
CT	0	1	0	12	0	0	0	3	0	16	75%
IC	0	0	0	0	6	0	0	0	0	6	100%
MI	0	0	0	0	0	9	0	0	4	13	69.23%
OS	0	0	0	0	0	0	7	0	0	7	100%
RU	0	0	9	0	0	0	0	0	0	9	0%
SR	0	0	13	0	0	0	0	0	0	13	0%
Total	13	13	39	13	6	9	7	11	4	115	

Table S3. Confusion matrix for *k*=9 for Hierarchical clustering showing precision values for all populations.

Population	1	4	2	5	7	8	3	6	9	Total	Precision
AR	13	0	4	0	0	0	0	0	0	17	76.47%
AT	0	17	0	2	0	0	0	2	0	21	80.95%
BS	0	0	13	0	0	0	0	0	0	13	100%
CT	0	1	0	16	0	0	0	0	0	16	100%
IC	0	0	0	0	6	0	0	0	0	6	100%
MI	0	0	0	0	0	9	0	0	4	13	69.23%
OS	0	0	0	0	0	0	7	0	0	7	100%
RU	0	0	9	0	0	0	0	0	0	9	0%
SR	0	0	13	0	0	0	0	0	0	13	0%
Total	13	17	39	18	6	9	7	2	4	115	

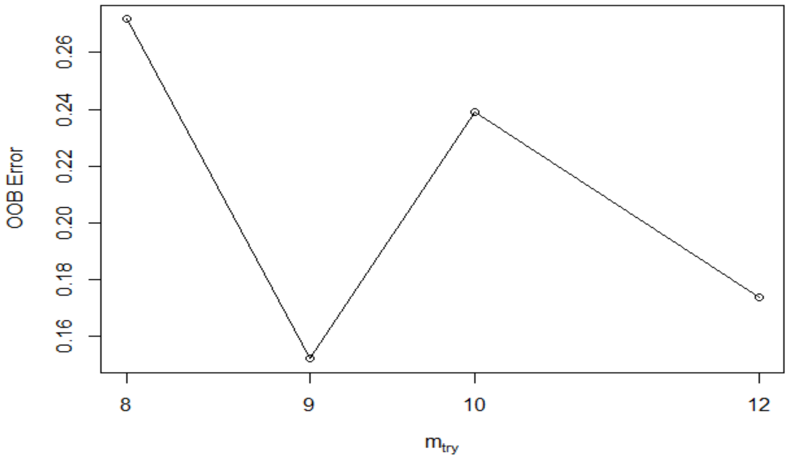

Figure S1. Mtry estimation using `tuneRF()` to find the value with the least OOB error.

Table S4. Confusion matrix for the trained random forest model using 20,000 trees and a mtry of 9 showing error rate per population.

	AR	AT	BS	CT	IC	MI	OS	RU	SR	Class Error
AR	12	0	0	0	0	0	0	0	3	20%
AT	0	12	0	2	0	0	0	0	0	14%
BS	0	0	7	0	0	0	0	1	2	30%
CT	0	2	0	10	0	0	0	0	0	17%
IC	0	0	0	0	5	0	0	0	0	0%
MI	0	0	0	0	0	9	0	0	0	0%
OS	0	0	0	0	0	0	7	0	0	0%
RU	0	0	3	0	0	0	0	5	0	38%
SR	1	0	1	0	0	0	0	1	9	25%