RESEARCH ARTICLE

Methods in Ecology and Evolution   BRITISH ECOLOGICAL SOCIETY

Using identity calls to detect structure in acoustic datasets

Taylor A. Hersh[1] iD | Shane Gero[1,2] iD | Luke Rendell[3] iD | Hal Whitehead[1] iD

[1]Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada

[2]Zoophysiology, Department of Bioscience, Aarhus University, Aarhus, Denmark

[3]Centre for Social Learning and Cognitive Evolution, School of Biology, University of St. Andrews, St. Andrews, UK

Correspondence
Taylor A. Hersh
Email: taylor.a.hersh@gmail.com

Abstract

1. Acoustic analyses can be powerful tools for illuminating structure within and between populations, especially for cryptic or difficult to access taxa. Acoustic repertoires are often compared using aggregate similarity measures across all calls of a particular type, but specific group identity calls may more clearly delineate structure in some taxa.

2. We present a new method—the identity call method—that estimates the number of acoustically distinct subdivisions in a set of repertoires and identifies call types that characterize those subdivisions. The method uses contaminated mixture models to identify call types, assigning each call a probability of belonging to each type. Repertoires are hierarchically clustered based on similarities in call type usage, producing a dendrogram with 'identity clades' of repertoires and the 'identity calls' that best characterize each clade. We validated this approach using acoustic data from sperm whales, grey-breasted wood-wrens and Australian field crickets, and ran a suite of tests to assess parameter sensitivity.

3. For all taxa, the method detected diagnostic signals (identity calls) and structure (identity clades; sperm whale subpopulations, wren subspecies and cricket species) that were consistent with past research. Some datasets were more sensitive to parameter variation than others, which may reflect real uncertainty or biological variability in the taxa examined. We recommend that users perform comparative analyses of different parameter combinations to determine which portions of the dendrogram warrant careful versus confident interpretation.

4. The presence of group-characteristic identity calls does not necessarily mean animals perceive them as such. Fine-scale experiments like playbacks are a key next step to understand call perception and function. This method can help inform such studies by identifying calls that may be salient to animals and are good candidates for investigation or playback stimuli. For cryptic or difficult to access taxa with group-specific calls, the identity call method can aid managers in quantifying behavioural diversity and/or identifying putative structure within and between populations, given that acoustic data can be inexpensive and minimally invasive to collect.

1 | INTRODUCTION

The natural world is inherently noisy, and animals that communicate acoustically must evolve ways of ensuring signal transmission fidelity in the face of such noise. Variation in acoustic signals can provide information about the identity of individuals, groups, populations and species, as well as drivers of signal evolution (Wilkins et al., 2013). Biologists who want to understand animal communication systems need robust, efficient ways of classifying and comparing acoustic signals.

Animal calls (i.e. acoustic signals) and repertoires (i.e. collections of calls) have often been compared using measures of similarity calculated over a range of call parameters. However, the presence of group-specific calls in some taxa (e.g. canids, Kershenbaum et al., 2016; sperm whales, Rendell & Whitehead, 2003) may indicate various levels of biological structure (e.g. social groups, cultures, subpopulations, populations, subspecies, species, etc.) to human observers more clearly than comparisons of aggregate similarity measures across calls or repertoires. This is especially likely if those group-specific calls—which we term 'identity calls'—have been produced by selection pressures for signalling identity, since signal detection theory predicts that calls with such functions will exhibit transmission-enhancing features, including redundancy, stereotypy and distinctiveness (Wiley, 2013).

We present a new method—the identity call method (hereafter IDcall)—that finds redundant, stereotyped and distinct identity calls in acoustic datasets and uses those identity calls to predict corresponding biological structure. Calls are first indexed in multivariate space using appropriate quantitative measures. Then, call types are identified by applying parsimonious mixtures of multivariate contaminated normal distributions (hereafter contaminated mixture models) to this multivariate dataset, with each call assigned a probability of belonging to each type. Lastly, repertoires (from single or multiple individuals, Method S2) are hierarchically clustered based on similarities in call type usage. The final dendrogram shows 'identity clades' (i.e. clades of repertoires distinguished by certain call types) and the identity calls that characterize each clade. We use 'clade' in a non-evolutionary sense to mean a node and all branches descended from it on a dendrogram.

We illustrate the efficacy of this method in finding identity calls that denote previously described biological structure using acoustic data from sperm whales *Physeter macrocephalus*, grey-breasted wood-wrens *Henicorhina leucophrys* (hereafter wrens) and Australian field crickets (*Teleogryllus* spp.; hereafter crickets). These datasets were selected to demonstrate the breadth of the method's capacity to detect different levels of biological structure —sperm whale subpopulations, wren subspecies and cricket species—using identity calls. In all cases, we knew biological structure existed a priori based on past research that often included extensive genetic and/or

mark–recapture analyses. Such approaches may not be appropriate or possible for taxa that are cryptic or difficult to access, and are often more costly and logistically challenging than acoustic data collection (Garland et al., 2015). We hope that IDcall can complement such approaches by allowing users to rapidly detect diagnostic signals (identity calls) and putative biological structure (identity clades) that may merit additional research to determine genetic, behavioural and/or ecological distinctiveness of the animals producing the repertoires, ultimately facilitating more targeted management and conservation action, if necessary.

2 | MATERIALS AND METHODS

2.1 | Classifying calls into types

IDcall was implemented in R version 3.6.1 (R Core Team, 2013). When applicable, the calls were first divided based on the number of quantitative measures (N). For example, sperm whale codas (i.e. stereotyped click patterns made in social situations) were divided based on the number of inter-click intervals. All calls, represented as points in N dimensional space, were classified into types using contaminated mixture models via the 'CONTAMINATEDMIXT' R package (Punzo et al., 2018). For each group of N-dimensional calls, we fitted between 2 and 10 mixture components to the data using the expectation conditional-maximization (ECM) algorithm initialized with the k-means algorithm (Punzo & McNicholas, 2016). We fit a range of mixture components to avoid a priori specification of the number of call types defined by the resulting components and used k-means based on the results of previous simulation studies (e.g. Shireman et al., 2017). To reduce the risk of outliers (i.e. calls that have ambiguous type) being assigned to their own mixture component(s) when fitting higher numbers of mixture components (e.g. Evans et al., 2015), we required all of the fitted models to have outliers by setting the 'contamination' Boolean to TRUE in the 'CNmixt' function (Punzo & McNicholas, 2016). The most parsimonious model and number of mixture components were selected based on the Bayesian information criterion (BIC) for large datasets (Schwarz, 1978) or the bias-corrected Akaike information criterion (AICc) for small datasets (Hurvich & Tsai, 1989). For each N, the final number of mixture components was the number of call types and the call types were validated by comparison to previous research (Method S3).

The contaminated mixture model algorithm estimates the probability that each call, i, belongs to each call type, j, as $u(i, j)$ (where $u(i, j) = 0$ if call i has a different N from the calls in j). We calculated usage, U, of each call type, j, for each repertoire, r, by summing the probability of call type membership for all calls in the repertoire and dividing by the total number of calls in the repertoire, $n(r)$:

$$U(r, j) = \frac{\sum_{i \in r} u(i, j)}{n(r)}.$$

A U close 0 means a call type is rarely used in the repertoire, whereas a U close to 1 means a call type is frequently used. If U equals 1, the call type is the only one used in the repertoire.

2.2 | Delineating identity calls and identity clades

Average linkage hierarchical clustering was used to cluster repertoires into clades based on quantitative similarity measures between pairs of repertoires, where similarity is calculated as the correlation of call type usages between each pair of repertoires (e.g. correlations of $U(r_1,:)$ with $U(r_2,:)$). Two repertoires that used the same call types with the same relative frequencies would have a high correlation value (close to 1) and cluster close together in the dendrogram, whereas two repertoires that used different call types and/or had opposite trends in call type usage would have a low correlation value (close to −1) and not cluster together. A cluster of repertoires could only be designated as an identity clade if those repertoires were united by high usage of at least one identity call—a call type used frequently by that clade and rarely by all others.

For a call type j to be considered an identity call for clade c, the mean call type usage across repertoires (mean $U(r \in c, j)$) in a clade had to be greater by a factor of at least a given value (a parameter we call *critfact*) than both the mean call type usage in each other clade with at least a given number of repertoires (a parameter we call *minrep*) at that level of the dendrogram and the mean call type usage of all repertoires not in that clade. Additionally, within a potential identity clade, a call type could only be designated as an identity call if the mean call type usage of the *minrep* repertoires in the clade that used the call type the most was less than *critfact* multiplied by the median call type usage across all repertoires in the clade. This prevented call types used frequently by just a few repertoires in a clade from being designated as identity calls. Extremely rare call types—for which the median usage in the *minrep* repertoires that used the call type the most was 0—were not used to build the dendrogram.

Starting at the base of the dendrogram (where each repertoire is a leaf), we ran an algorithm (Table 1) that tested each node in turn to see if it met the criteria to be considered an identity clade (i.e. a clade of at least *minrep* repertoires with at least one identity call in the final tree). In each identity clade, c, we calculated the proportion of repertoire calls that were identity calls (the clade identity call proportion, icp) from:

$$icp(c) = \frac{\sum_{r \in clade(c)} \sum_{j: \text{ identity call for clade}(c)} U(r, j)}{\sum_{r \in clade(c)} \sum_j U(r, j)}$$

To quantify the support for each repertoire belonging to an identity clade, we calculated the correlation between the call type usages of the repertoire and the median usages of the identity clade.

TABLE 1 Hierarchical clustering algorithm for daughter and parental clades. Each clade has one of three possible statuses: potential identity clade (P), final identity clade (F) or no classification (N; no additional classification occurs). Initially, each repertoire (considered a daughter clade) has status P. Clades with status P can be P− (no identity calls) or P+ (with identity calls; x and y denote different identity call types). Statuses of daughter and parental clades (the clades being merged and the resultant clade, respectively) change at each node following the rules below. When all clades have status F or N, or the correlation between two clades joined at a node is less than zero, the process terminates

Daughter clade 1 status	Daughter clade 2 status	Result
P−	P+	Parental clade becomes P+ (has identity calls) or P− (does not have identity calls)
P−	P−	
P+$_x$	P+$_y$	Both daughter clades become F and are not merged; parental clade is N
F	P+	Daughter clade that was P+ becomes F; parental clade is N
N	P+	
F	P−	Daughter clade that was P− becomes N; parental clade is N
N	P−	
F	F	Parental clade is N
N	N	
F	N	

Proximity of identity clades in the dendrograms reflects repertoire similarity but may or may not reflect phylogeny. Two neighbouring identity clades generally have repertoires that are more similar to each other than to other clades further away, but that does not necessarily mean they are descended from the same 'ancestor repertoire'. An understanding of how the calls comprising repertoires change over time is necessary before phylogenetic inferences can be attempted.

2.3 | Test datasets

We used acoustic data from sperm whales, wrens and crickets to test the ability of IDcall to find previously described biological structure. Our analyses are illustrative but should not be regarded as definitive results for each taxon.

Globally, sperm whale populations are divided by the geography of the different oceans, where only males move genes between basins (Lyrholm et al., 1999), but can also be delineated locally into sympatric cultural clans which are defined based on the codas they use (Rendell & Whitehead, 2003). Clans represent socially segregated subpopulations that are phenotypically distinct (Rendell & Whitehead, 2003). The sperm whale dataset contains 13,805 codas recorded in the Atlantic/Mediterranean (Table S1; Figure S1) and 18,481 codas recorded in the Pacific (Table S2; Figure S1). Our objective was to determine whether different clans could be distinguished

by 'identity codas' using IDcall applied to coda absolute inter-click intervals ($N = 2$–8). Codas were divided into repertoires based on the group identity of the recorded whales (Method S2).

The wren dataset (Halfwerk et al., 2016; Dryad Digital Repository https://doi.org/10.5061/dryad.q5p7g) contains 396 averaged song types from males of two subspecies (*H. l. hilaris* and *H. l. leucophrys*) whose ranges overlap but eventually replace each other with changing altitude in the Ecuadorian Andes. These subspecies are genetically distinct but morphologically similar and can be delineated based on their characteristic song (Dingle et al., 2008). Our goal was to determine whether IDcall could distinguish 'identity songs' of the two subspecies using several song frequency parameters (averaged note peak frequency, minimum and maximum song frequency; $N = 3$; Method S2). Songs were divided into repertoires by individual ($n = 41$ males; Method S2).

The cricket dataset (Moran et al., 2020; Dryad Digital Repository https://doi.org/10.5061/dryad.wpzgmsbhr) is comprised of male calling song data for two closely related species (*Teleogryllus commodus*, 127 songs; *Teleogryllus oceanicus*, 131 songs) which live in sympatry across hundreds of kilometres of the Australian east coast and show similar habitat and resource use (Moran et al., 2020). The two species produce hybrids in the laboratory but rarely or not at all in the wild, with acoustics helping prevent wild crossbreeding (Moran et al., 2020). Our objective was to see whether IDcall could distinguish *T. commodus* and *T. oceanicus* 'identity songs' using several interval-based song features (chirp pulse length, chirp interpulse interval, chirp-trill interval and trill-pulse length; $N = 4$; Method S2). Songs were divided into repertoires by field site ($n = 16$ sites; Method S2).

2.4 | Testing options/parameters and comparing dendrograms

Several options and parameters must be set in IDcall. To assess how varying each impacted the final identity calls and clades, we tested a range of settings/values, changing one at a time while keeping the others at established defaults (Table 2), and compared each 'trial dendrogram' to a 'baseline dendrogram'. The default values should not be interpreted as optimized values for each dataset but provided a reasonable starting point to assess variation across trials; other settings may be more appropriate for other datasets (see Section 4.5 for recommendations). To assess start-point dependence during call classification, we reran the default parameters twice and compared both dendrograms to the baseline dendrogram.

Within and across datasets, and for each trial ($n = 19$ per dataset), we assessed how the number of call types (total and identity), number of identity clades and the tree identity call proportion varied. We also compared the repertoire composition of identity clades in each trial to the baseline by looking at all possible repertoire pairs and determining if each pair was assigned to the same or different clade in the two dendrograms. All repertoires not assigned to an identity clade were considered part of a single 'outlier clade'. Similarity was

calculated as the number of repertoire pairs assigned to the same clade in both dendrograms plus the number of pairs assigned to different clades in both dendrograms divided by the total number of pairs. To verify that the method was not prone to detecting spurious identity calls and clades, we randomly assigned calls to repertoires 10 times for each dataset and used the default values to see whether identity calls and clades were still delineated.

3 | RESULTS

3.1 | General method performance

For all four datasets, IDcall found identity calls and delineated identity clades (for baseline dendrogram results, see Table S3). The Atlantic/Mediterranean sperm whale, wren and cricket dendrograms were robust (in terms of number and composition of identity clades) across trials, while the Pacific sperm whale dendrograms exhibited more variability (Tables S4–S7, Supplemental Data). When the repertoire grouping variable of each call was randomly permuted, no identity calls or clades were delineated in any dataset.

As *critfact* increased, the number of identity calls and tree identity call proportion generally decreased. The number of identity clades did not change across trials for Atlantic/Mediterranean sperm whales or crickets, but gradually decreased for Pacific sperm whales and decreased at the highest tested value for wrens.

With increasing *minrep*, the number of identity clades generally decreased. For the sperm whale and cricket datasets, the number of identity calls decreased as well. The opposite trend was seen for wrens. The tree identity call proportion decreased for Pacific sperm whales and crickets but increased for Atlantic/Mediterranean sperm whales and wrens. When *minrep* was increased beyond the expected identity clade size, no identity clades were delineated.

Using random.post (i.e. random generation of the initial matrix with posterior probabilities of group membership; Punzo et al., 2018) for ECM initialization generally increased the number of call types (total and identity) compared to the baseline for sperm whales but had the opposite effect for wrens/crickets. With random.clas (i.e. random generation of the initial classification matrix; Punzo et al., 2018), the effect on the total number of call types varied across datasets, but the number of identity calls and tree identity call proportion decreased for the Atlantic/Mediterranean sperm whale, wren and cricket datasets (the opposite was seen for Pacific sperm whales). Under both strategies, the number and composition of identity clades was similar to the baseline except for Pacific sperm whales when using random.post (a baseline dendrogram identity clade was split in two) and for wrens when using random.clas (no identity clades delineated).

Across datasets, using AIC during call classification resulted in the most call types while ICL resulted in the fewest. Using AIC typically decreased the number of identity calls from the baseline criterion (BIC for sperm whales, AICc for wrens/crickets), but the effect on tree identity call proportion varied across datasets and

TABLE 2 Default and trial values for each IDcall option and parameter. Default settings were used across datasets unless noted. In the wren and cricket datasets, slightly smaller *minrep* values were tested and AICc was used as the default information criterion to account for the smaller dataset sizes

Stage	Option/parameter	Default setting	Additional tested settings/values	General method behaviour
Call classification	ECM initialization strategy	*k*-means	random.post, random. clas	Effect on number of call types (total and identity) and tree identity call proportion varied across datasets but typically yielded similar trees
	Information criterion	BIC (sperm whales) AICc (crickets, wrens)	AIC, ICL, AICc (sperm whales) AIC, ICL, BIC (crickets, wrens)	Using AIC/AICc produced the most call types. Using ICL produced the least, with BIC intermediate
Hierarchical clustering	Linkage method	average	single, complete	Single linkage resulted in chained trees with long, straggly clusters. Complete linkage resulted in trees with compact clusters but more repertoires designated as outliers. Average linkage trees were more similar to single linkage trees than to complete linkage trees
	critfact	14	6, 10, 18, 22, 26	Increasing *critfact* corresponded with a decreasing number of identity calls/clades and tree identity call proportion
	minrep	6	4, 8, 10, 12, 14 (sperm whales) 3, 5, 7, 9, 11 (crickets, wrens)	Increasing *minrep* corresponded with a decreasing number of identity calls/clades. Effect on tree identity call proportion varied across datasets

the final number of identity clades did not change in any dataset. In contrast, using ICL decreased the number of identity calls and the tree identity call proportion, with the number of identity clades remaining constant (sperm whales) or decreasing (wrens and crickets). Start-point dependence during call classification was evidenced by variation in the number of call types (total and identity) in duplicate runs of the default parameters across datasets, but the final identity clades were very similar to those in the baseline dendrograms.

Using single linkage during hierarchical cluster analysis resulted in chained trees with long, straggly clusters, whereas complete linkage produced trees with compact clusters but more repertoires designated as outliers (or no tree at all for wrens). Average linkage and single linkage trees were generally more similar to each other than to complete linkage trees.

3.2 | Sperm whales

For most parameter combinations ($n = 16/20$, including the baseline), the Atlantic/Mediterranean sperm whale coda repertoires were delineated into three identity clades corresponding to the two known eastern Caribbean clans (EC1 and EC2; Gero et al., 2016) and the known Mediterranean clan (Rendell & Frantzis, 2016) (for baseline dendrogram, see Figure 1 and Figure S2; average trial dendrogram similarity = 0.992 ± 0.017). The Gulf of Mexico and Panama repertoires were outliers. The EC2 clan had the most identity calls ($n = 5$), followed by the Mediterranean ($n = 3$) and EC1 ($n = 2$) (Table S8). Only the EC1 and Mediterranean clans were detected at higher

values of *minrep*. In the complete linkage dendrogram, all EC1 repertoires were designated as outliers.

The Pacific sperm whale baseline tree had five identity clades (Figure 2; Figure S3), four of which correspond to known clans (Regular, Short, Plus-One and Four-Plus). The putative fifth clan includes repertoires previously considered Four-Plus (Cantor et al., 2016) and a new repertoire. The number of identity codas (2–7) and the clan identity coda proportion (22%–56%) varied across clans (Table S9). Results across trials for Pacific sperm whales were more variable, which was reflected in a lower average trial dendrogram similarity (0.972 ± 0.037) than Atlantic/Mediterranean sperm whales. Across trials, the Regular ($n = 20/20$), Plus-One ($n = 19/20$), Four-Plus ($n = 20/20$) and putative fifth ($n = 16/20$) clans were delineated robustly except when *minrep* exceeded the number of repertoires in the clan. The most variability was seen in the baseline Short clan, which often formed one clan ($n = 13/20$) but sometimes two ($n = 4/20$, e.g. Figure S4) or three ($n = 1/20$, e.g. Figure S5) depending on the trial. The complete linkage dendrogram looked the most distinct from the baseline, with many outlier repertoires ($n = 34$). In the single linkage dendrogram, a smaller version of the Short clan formed, with many of the baseline Short clan repertoires assigned to the Four-Plus clan.

3.3 | Wrens

Identity clades were typically delineated in the wren dataset ($n = 17/20$; no identity clades when using random.clas or BIC/ICL) and, in most instances ($n = 14/17$), they matched the two subspecies: *H. l.*
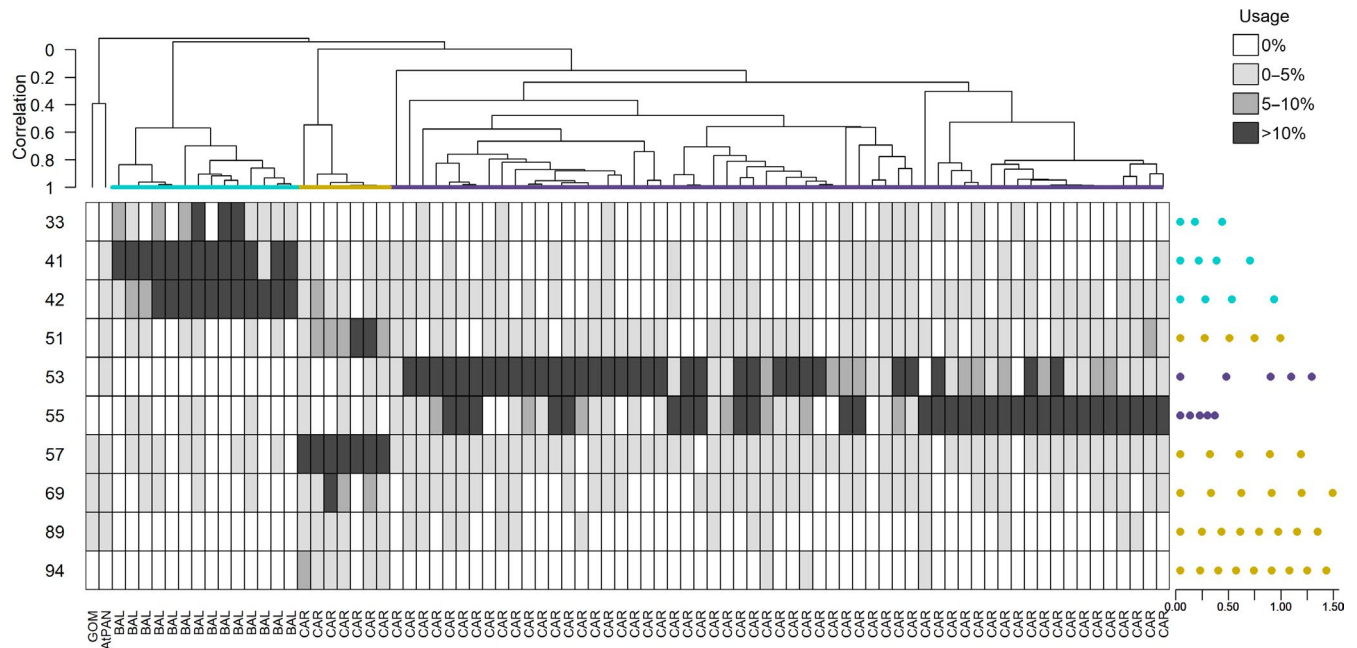
FIGURE 1 Baseline dendrogram with clans and identity coda types for Atlantic/Mediterranean sperm whales. Average linkage hierarchical clustering dendrogram (top) depicts similarity among sperm whale coda repertoires recorded in the Atlantic/Mediterranean. Coloured identity clades correspond to three known clans: Mediterranean (cyan), EC2 (gold), and EC1 (purple). Heat map (bottom) depicts identity coda type usage (rows) for each repertoire (columns) in shades of grey, which differs from similar figures in past sperm whale studies (e.g. Cantor et al., 2016) because usage here is calculated based on probabilistic assignment of codas to types. Identity coda type codes are on the left (see Table S12 for type names) and type centroid rhythm plots (coloured by clan) are on the right (each dot represents a click; scale bar is in seconds). Recording locations are listed on the bottom (see Figure S1 for abbreviations)
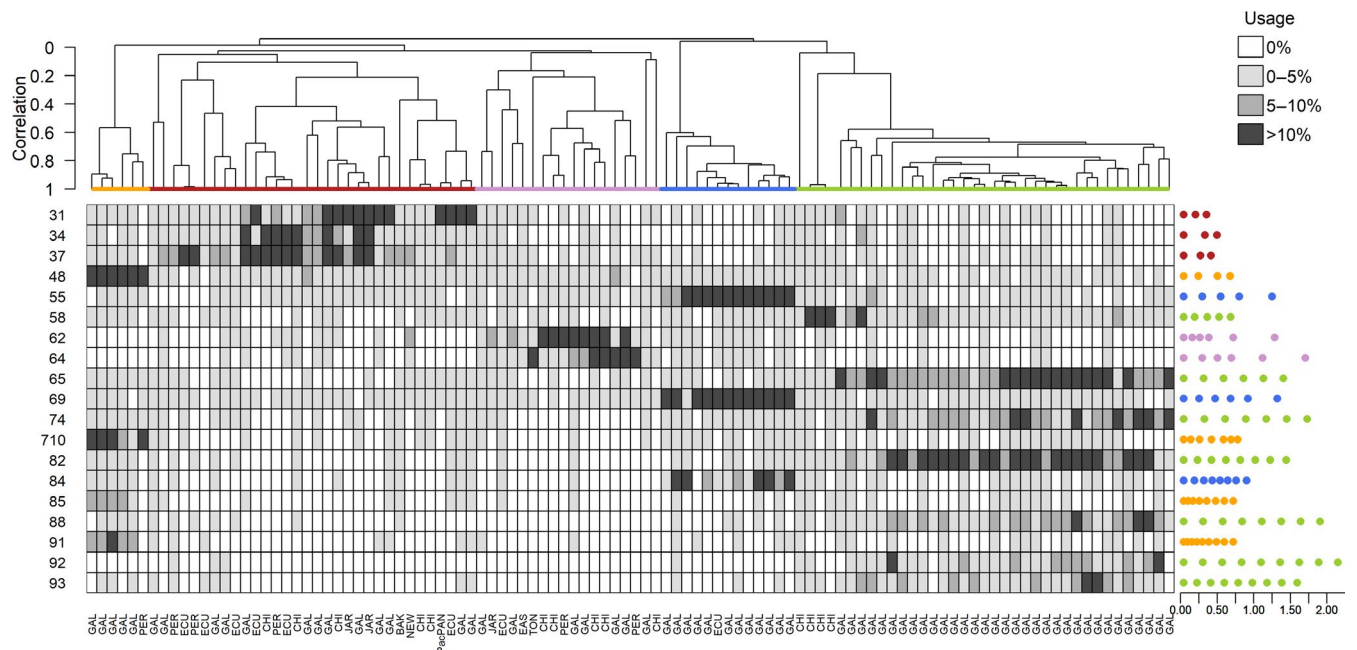

FIGURE 2 Baseline dendrogram with clans and identity coda types for Pacific sperm whales. Average linkage hierarchical clustering dendrogram (top) depicts similarity among sperm whale coda repertoires recorded in the Pacific. Coloured identity clades correspond to a putative new clan (orange) and four known clans: Short (red), Four-Plus (pink), Plus-One (blue), and Regular (green). Heat map (bottom) depicts probabilistic identity coda type usage (rows) for each repertoire (columns) in shades of grey, which differs from similar figures in past sperm whale studies (e.g. Cantor et al., 2016) because usage here is calculated based on probabilistic assignment of codas to types. Identity coda type codes are on the left (see Table S13 for type names) and type centroid rhythm plots (coloured by clan) are on the right (each dot represents a click; scale bar is in seconds). Recording locations are listed on the bottom (see Figure S1 for abbreviations)

hilaris (12 males) and *H. l. leucophrys* (29 males) (Figure 3; Figure S6; average trial dendrogram similarity = 0.989 ± 0.037). *H. l. hilaris* had one identity song and *H. l. leucophrys* had three (Table S10). Repertoires of four F1 hybrid birds clustered with *H. l. leucophrys* while one clustered with *H. l. hilaris*. The six second-generation males clustered according to their parental subspecies and two *H. l. hilaris* males clustered with *H. l. leucophrys*. The identity songs for the two subspecies differ in averaged note peak frequency and minimum/maximum song frequency, with the *H. l. hilaris* values lower than the *H. l. leucophrys* values. Using random.post as the ECM initialization, two additional genetically *H. l. hilaris* birds clustered with *H. l. leucophrys*. Only the *H. l. hilaris* clade was detected using complete linkage and only the *H. l. leucophrys* clade was detected at the highest value of *critfact*.

3.4 | Crickets

In most trials (*n* = 16/20), identity clades corresponding to the two species, *T. oceanicus* and *T. commodus*, were detected in the cricket dataset (for baseline dendrogram, see Figure 4 and Figure S7; average trial dendrogram similarity = 1.00 ± 0.00). The clade identity song proportion was similar and high for the *T. oceanicus* sites (99%, two identity songs) and the *T. commodus* sites (98%, two identity songs) (Table S11). The centroid chirp-trill and trill pulse lengths were

shorter in *T. commodus* identity songs than in *T. oceanicus* identity songs. No identity clades were produced at the two highest tested values of *minrep* and only a single identity clade (*T. commodus*) was produced using BIC/ICL.

4 | DISCUSSION

4.1 | General method performance

Contaminated mixture modelling has several strengths as a classification method. It (a) minimizes the number of parameters specified a priori; (b) identifies outliers, letting the user decide if the outliers should be retained (as we did in this study) or excluded; (c) allows clusters to have varying volume, shape and orientation in multivariate space; and (d) can be used on both low- and high-dimensional data (Punzo & McNicholas, 2016). Capitalizing on these advantages, our work demonstrates the versatility of this classification method.

The choice of how to divide calls into repertoires should be guided by features of the user's dataset (e.g. sampling resolution) and research question (e.g. looking for individual versus group identity signals), as it was here (Method S2). When calls were randomly permuted among repertoires, IDcall did not delineate identity calls or clades in the dendrogram for any dataset for any
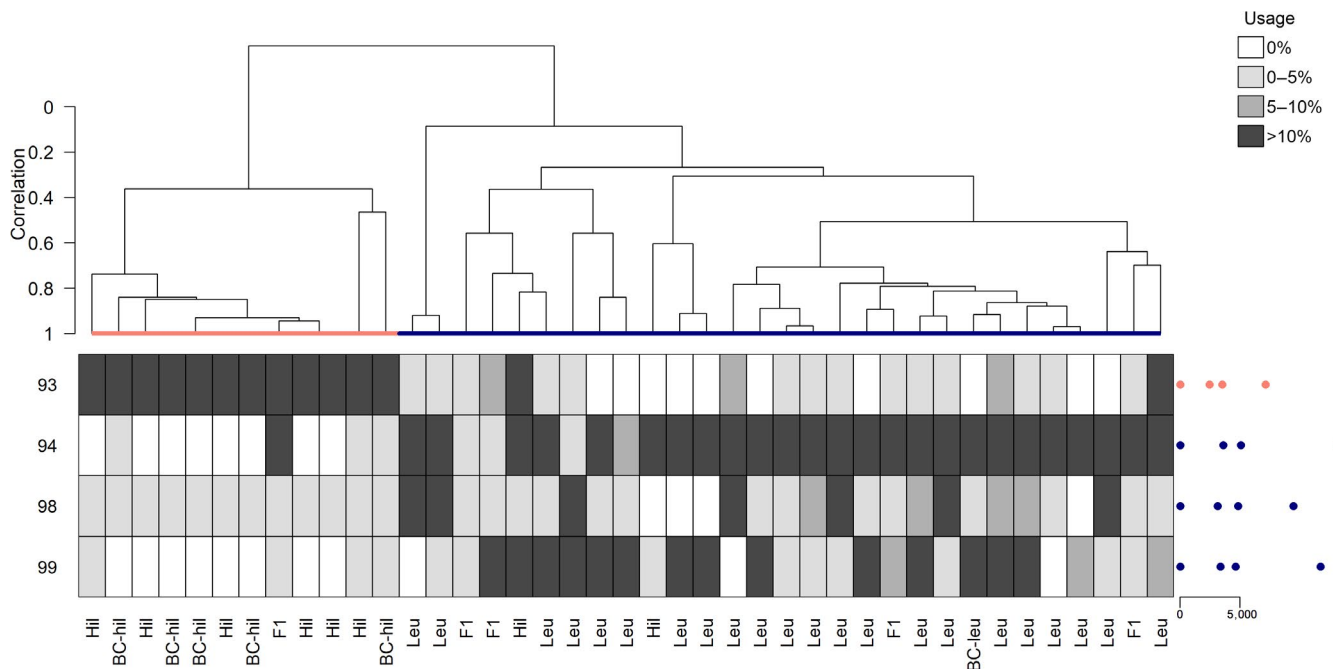

FIGURE 3 Baseline dendrogram with subspecies and identity song types for wrens. Average linkage hierarchical clustering dendrogram (top) depicts similarity among song frequency vectors of male wrens. Coloured identity clades correspond to two subspecies: *H. l. hilaris* (salmon) and *H. l. leucophrys* (navy). Heat map (bottom) depicts identity song type usage (rows) for each male (columns) in shades of grey, with usage calculated based on probabilistic assignment of songs to types. Identity song codes are on the left and frequency centroid vector plots are on the right. From left to right, the space between the dots represents averaged note peak frequency, minimum song frequency, and maximum song frequency (scale bar is in Hz) for each song type. Clustering was done on logged frequency vectors, but vector plots are presented unlogged to aid in interpretation. Genotyping abbreviations are: Hil, parental *H. l. hilaris*; Leu, parental *H. l. leucophrys*; F1, first-generation hybrid; BC-hil, backcross between Hil and F1; and BC-leu, backcross between Leu and F1 (Halfwerk et al., 2016)
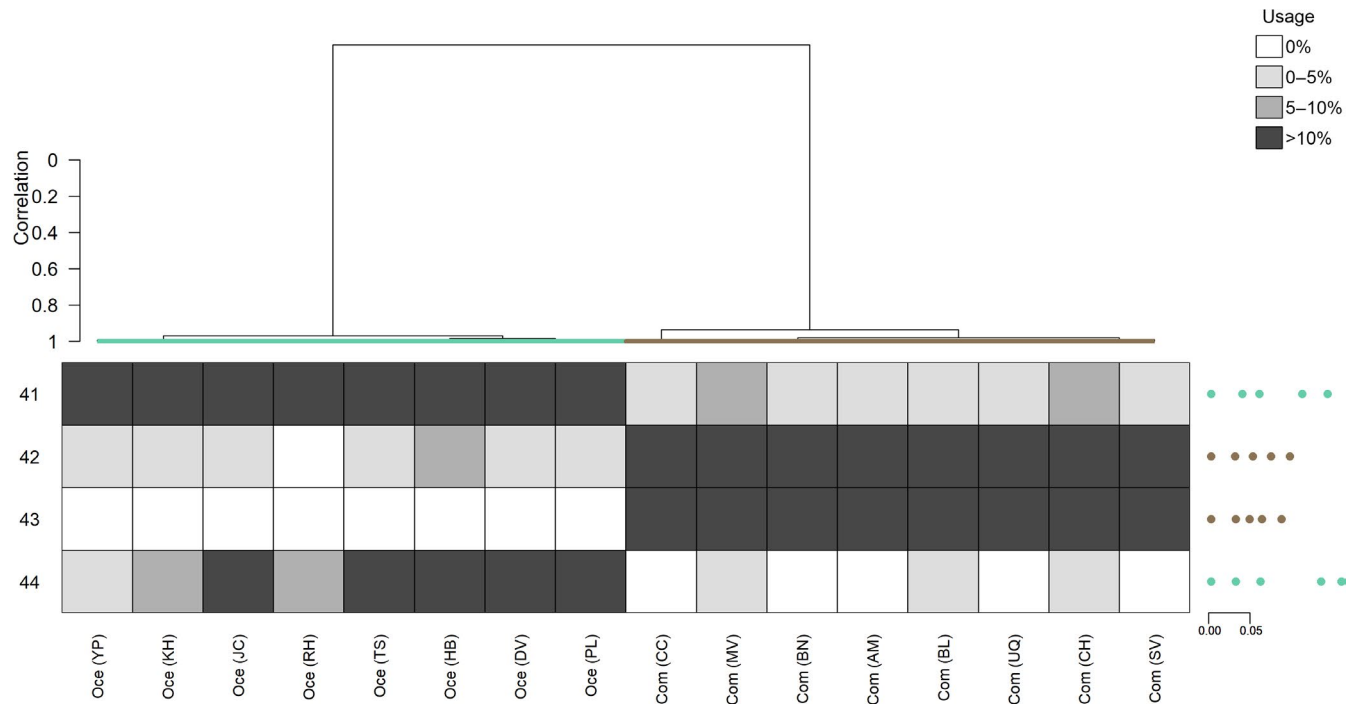
FIGURE 4 Baseline dendrogram with species and identity song types for crickets. Average linkage hierarchical clustering dendrogram (top) depicts similarity among song interval vectors of male crickets from 16 sites. Coloured identity clades correspond to two species: *Teleogryllus oceanicus* (teal) and *Teleogryllus commodus* (brown). Heat map (bottom) depicts identity song type usage (rows) for each field site (columns) in shades of grey, with usage calculated based on probabilistic assignment of songs to types. Identity song codes are on the left and interval centroid vector plots are on the right. From left to right, the spaces between the dots represent chirp pulse length, chirp interpulse interval, chirp-trill interval, and trill pulse length (scale bar is in seconds). Species abbreviations (Com, *T. commodus*; Oce, *T. oceanicus*) are listed along the bottom, with corresponding field site names in parentheses (Moran et al., 2020)

permutation (*n* = 40 total). Call types that were characteristic of a given repertoire (and may have eventually become identity calls for a clade) were instead scrambled across repertoires. This suggests the method is unlikely to delineate identity calls and clades when they do not exist. In such cases, a dendrogram is still produced but it does not have identity calls or clades. However, the method could also struggle to detect true identity calls and clades for small datasets. For example, when the number of repertoires in the Atlantic/Mediterranean sperm whale dataset was randomly halved, the EC1 (*n* = 31 repertoires) and Mediterranean (*n* = 7) vocal clans were still detected using the default parameters (Figure S8) but the EC2 clan (*n* = 3) was not detected until *minrep* was decreased to 3 (Figure S9). When the number of repertoires was randomly quartered, no clans were delineated using the default parameters, but the EC1 (*n* = 14) and Mediterranean (*n* = 5) vocal clans were delineated when *minrep* was decreased to 5 (Figure S10). The EC2 vocal clan had only one repertoire in this randomized subset, and never became a clan. As these examples illustrate, the method's performance relies on the size and features of the input data.

During call classification, using AIC/AICc generally resulted in the highest total number of call types, ICL the lowest and BIC intermediate. Using complete linkage during tree building produced the most distinct trees, whereas average and single linkage behaved more similarly. Increasing *critfact* restricted the requirements to be

considered an identity call, so fewer calls were identified as such, and increasing *minrep* prohibited small identity clades with few repertoires from forming. Varying the ECM algorithm initialization strategy affected the number of calls (total and identity) and tree identity call proportion, but the direction of these effects varied by dataset. Aside from the wrens (for which random.post misassigned two birds and random.clas did not produce any identity clades), the three strategies generally yielded similar results.

4.2 | Sperm whales

Most of the sperm whale clans detected by IDcall agree with past work (Method S3; Tables S12 and S13) and many of the identity codas have previously been recognized as indicators of clan identity (Cantor et al., 2016; Gero et al., 2016; Pavan et al., 2000; Rendell & Whitehead, 2003). Coda type usage results from recent sperm whale studies suggest that the clan identity coda phenomenon extends beyond the locations considered here (e.g. Brazil: Amorim et al., 2020; Mauritius: Huijser et al., 2020).

The Atlantic/Mediterranean sperm whale dataset was reliably delineated into three clans across trials except when *minrep* exceeded the number of EC2 clan repertoires. The number and composition of Pacific sperm whale clans was more variable, but this variation centred on the Short clan. The robustness of the other

clans (Regular, Plus-One, Four-Plus and putative fifth) across trials lends confidence to those clan designations. While the repeated segregation of the putative fifth clan from the Four-Plus clan across trials could suggest that IDcall is not performing as well as previous methods (e.g. Cantor et al., 2016), the differences in identity call usage in both identity clades, coupled with no photo-identified groups of whales linking any of the Four-Plus or putative fifth clan repertoires, support them as separate clans. More data and analyses are necessary, but the current results suggest that IDcall is sensitive enough to identify clans that were not detected with other methods.

The repertoires comprising the baseline dendrogram Short clan typically clustered together across trials, but formed one to three clans (Figure 2; Figures S4 and S5) depending on the parameters. In the past, the Short clan has been regarded as an anomaly compared to other Pacific clans, given that the clan's most common coda types (e.g. $1 + 2$, $3R$, $2 + 1$) do not follow a characteristic rhythmic pattern across click lengths like other clans (Rendell & Whitehead, 2003). Indeed, different combinations of $1 + 2$, $3R$ and $2 + 1$ coda types become identity codas for the variations of the Short clan found using IDcall. Short clan coda type usage plots from past work (e.g. Cantor et al., 2016; Rendell & Whitehead, 2003) show that different repertoire subgroups within the clan vary in how much they use the aforementioned coda types, and each photo-identified group of whales contributed only one repertoire to the present analysis. This suggests that what has previously been referred to as the Short clan could be one or several clans, but more acoustic and photographic data are needed to tease these possibilities apart.

4.3 | Wrens

Using frequency measures, IDcall accurately clustered songs from male wrens into two subspecies. Consistent with the original study (Halfwerk et al., 2016), two *H. l. hilaris* males clustered with *H. l. leucophrys* in the baseline dendrogram. The distribution of F1 hybrid birds in the baseline dendrogram aligns with the original study as well (Halfwerk et al., 2016). The identity songs were consistent across trials and emphasized known song difference (Method S3; Halfwerk et al., 2016). That only the *H. l. leucophrys* clade was detected at the highest level of *critfact* suggests that *H. l. leucophrys* birds use their 'identity songs' more than *H. l. hilaris* birds.

4.4 | Crickets

The cricket dataset was robust to parameter variation. In most trials, IDcall accurately clustered male cricket calling songs into two species, and the detected identity songs emphasized known song differences (Method S3; Moran et al., 2020). Unsurprisingly, the two species were not delineated when *minrep* was increased beyond the number of repertoires for each species ($n = 8$). While no wren

identity clades and only one of the two expected cricket identity clades (*T. commodus*) were detected when using BIC/ICL, both datasets are small and AICc is likely the most appropriate choice for information criterion (Hurvich & Tsai, 1989).

4.5 | Method recommendations, limitations and applications

The IDcall options/parameters that were varied during our trials can be set by the user and informed by features of their dataset and research question. Using *k*-means as the ECM initialization strategy is effective when there are numerous local optima, and likely represents a good starting point for most users, but random.post and random.clas can be more computationally efficient (Shireman et al., 2017). Prior to call classification, users should examine their raw call data to get a better sense for how separated call clusters are. The number of mixture components (which BIC aims to find) may not always equal the number of clusters (which ICL aims to find), which can lead to different results depending on the dataset (Baudry et al., 2010). Generally, we recommend using BIC for large datasets with poorly separated clusters (Schwarz, 1978); ICL for large datasets with well-separated clusters (Biernacki et al., 2000) and AICc for small datasets (Hurvich & Tsai, 1989). Users can also set the range of number of mixture components to fit to the data during call classification.

Average linkage represents a middle ground between the extremes of single linkage and complete linkage and is likely an appropriate linkage method for most datasets. Using high values of *critfact* will generally result in fewer but more conservative identity calls (i.e. identity calls that are used very frequently by their identity clade and very rarely by all other clades) and using high values of *minrep* will typically result in fewer, larger identity clades. Users can thus vary *critfact* and *minrep* based on their desired level of conservativeness at both the identity call and identity clade level. The comprehensive output also lets users control metric cut-offs like repertoire/clade correlation. Repertoires that have low call type usage correlation with the rest of an identity clade can be excluded from subsequent analyses.

IDcall does have limitations. It will not cluster calls represented by different vector lengths (i.e. occupying multivariate spaces with varying dimensionality, *N*) at the same time. It also requires several parameters to be set, and these settings can impact the final dendrograms, as evidenced by some of the trial results. We recommend that users test different parameter combinations and compare the dendrograms. Dendrogram regions that are more sensitive warrant careful interpretation but may reflect underlying uncertainty in the data and/or true biological variability, whereas regions that are robust can likely be interpreted with more confidence.

We make no claims that IDcall will work for all acoustic animals, but its ability to find and use identity calls to detect structure—including subpopulations, subspecies and species—in three diverse taxa is promising. However, the presence of biologically informative

identity calls does not necessarily mean those calls are meaningful to the animals themselves. Playback experiments are key to test whether specific calls are used by animals to broadcast or infer identity. IDcall can inform such studies by identifying calls that are potentially characteristic of underlying biological structure (and likely to be perceptually salient) and are thus good targets for additional research or to be played back. For animals that are cryptic or difficult to access but have identity calls, IDcall can potentially aid managers in quantifying behavioural diversity and/or identifying putative discrete units, given that acoustic data can be inexpensive and minimally invasive to collect. This is particularly true in marine environments, where passive acoustic recording is relatively easy, whereas genetic or photographic sampling is often logistically complex.

AUTHORS' CONTRIBUTIONS

T.A.H. and H.W. conceived and designed the method, with input from S.G. and L.R.; T.A.H. analysed and interpreted the data and drafted the manuscript. All authors collected sperm whale coda data, revised the manuscript, and read and approved the final version.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13644.

DATA AVAILABILITY STATEMENT

The IDcall code is under active development by the authors and can be accessed, along with the sperm whale datasets, through the Open Science Framework (https://osf.io/5fter/). We welcome feedback on ways to improve and expand the method. The wren (Halfwerk et al., 2016) and cricket (Moran et al., 2020) data can be accessed through Dryad at https://doi.org/10.5061/dryad.q5p7g and https://doi.org/10.5061/dryad.wpzgmsbhr, respectively.

ORCID

Taylor A. Hersh 🔟 https://orcid.org/0000-0002-7891-596X
Shane Gero 🔟 https://orcid.org/0000-0001-6854-044X
Luke Rendell 🔟 https://orcid.org/0000-0002-1121-9142
Hal Whitehead 🔟 https://orcid.org/0000-0001-5469-3429

REFERENCES

Amorim, T. O. S., Rendell, L., Di Tullio, J., Secchi, E. R., Castro, F. R., & Andriolo, A. (2020). Coda repertoire and vocal clans of sperm whales in the western Atlantic Ocean. *Deep-Sea Research Part I: Oceanographic Research Papers*, *160*, 103254. https://doi.org/10.1016/j.dsr.2020.103254

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, *19*(2), 332–353. https://doi.org/10.1198/jcgs.2010.08111

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the Integrated Classification Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725. https://doi.org/10.1109/34.865189

Cantor, M., Whitehead, H., Gero, S., & Rendell, L. (2016). Cultural turnover among Galápagos sperm whales. *Royal Society Open Science*, *3*(10), 160615. https://doi.org/10.1098/rsos.160615

Dingle, C., Halfwerk, W., & Slabbekoorn, H. (2008). Habitat-dependent song divergence at subspecies level in the grey-breasted wood-wren. *Journal of Evolutionary Biology*, *21*(4), 1079–1089. https://doi.org.10.1111/j.1420-9101.2008.01536.x

Evans, K., Love, T., & Thurston, S. W. (2015). Outlier identification in model-based cluster analysis. *Journal of Classification*, *32*(1), 63–84. https://doi.org/10.1007/s00357-015-9171-5

Garland, E. C., Goldizen, A. W., Lilley, M. S., Rekdahl, M. L., Garrigue, C., Constantine, R., Hauser, N. D., Poole, M. M., Robbins, J., & Noad, M. J. (2015). Population structure of humpback whales in the western and central South Pacific Ocean as determined by vocal exchange among populations. *Conservation Biology*, *29*(4), 1198–1207. https://doi.org/10.1111/cobi.12492

Gero, S., Bøttcher, A., Whitehead, H., & Madsen, P. T. (2016). Socially segregated, sympatric sperm whale clans in the Atlantic Ocean. *Royal Society Open Science*, *3*(6), 160061. https://doi.org/10.1098/rsos.160061

Halfwerk, W., Dingle, C., Brinkhuizen, D. M., Poelstra, J. W., Komdeur, J., & Slabbekoorn, H. (2016). Sharp acoustic boundaries across an altitudinal avian hybrid zone despite asymmetric introgression. *Journal of Evolutionary Biology*, *29*(7), 1356–1367. https://doi.org/10.1111/jeb.12876

Huijser, L. A. E., Estrade, V., Webster, I., Mouysset, L., Cadinouche, A., & Dulau-Drouot, V. (2020). Vocal repertoires and insights into social structure of sperm whales (*Physeter macrocephalus*) in Mauritius, southwestern Indian Ocean. *Marine Mammal Science*, *36*(2), 638–657. https://doi.org/10.1111/mms.12673

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

Kershenbaum, A., Root-Gutteridge, H., Habib, B., Koler-Matznick, J., Mitchell, B., Palacios, V., & Waller, S. (2016). Disentangling canid howls across multiple species and subspecies: Structure in a complex communication channel. *Behavioural Processes*, *124*, 149–157. https://doi.org/10.1016/j.beproc.2016.01.006

Lyrholm, T., Leimar, O., Johanneson, B., & Gyllensten, U. (1999). Sex-biased dispersal in sperm whales: Contrasting mitochondrial and nuclear genetic structure of global populations. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*(1417), 347–354. https://doi.org/10.1098/rspb.1999.0644

Moran, P. A., Hunt, J., Mitchell, C., Ritchie, M. G., & Bailey, N. W. (2020). Sexual selection and population divergence III: Interspecific and intraspecific variation in mating signals. *Journal of Evolutionary Biology*, *33*(7), 990–1005. https://doi.org/10.1111/jeb.13631

Pavan, G., Hayward, T. J., Borsani, J. F., Priano, M., Manghi, M., Fossati, C., & Gordon, J. (2000). Time patterns of sperm whale codas recorded in the Mediterranean Sea 1985–1996. *The Journal of the Acoustical*

Society of America, *107*(6), 3487–3495. https://doi.org/10.1121/1.429419

Punzo, A., Mazza, A., & McNicholas, P. D. (2018). ContaminatedMixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. *Journal of Statistical Software*, *85*(10), 1–25. https://doi.org/10.18637/jss.v085.i10

Punzo, A., & McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, *58*(6), 1506–1537. https://doi.org/10.1002/bimj.201500144

R Core Team. (2013). *R: A language and environment for statistical computing*. R Development Core Team.

Rendell, L., & Frantzis, A. (2016). Mediterranean sperm whales, *Physeter macrocephalus*: The precarious state of a lost tribe. *Advances in Marine Biology*, *75*, 37–74. https://doi.org/10.1016/bs.amb.2016.08.001

Rendell, L. E., & Whitehead, H. (2003). Vocal clans in sperm whales (*Physeter macrocephalus*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, *270*(1512), 225–231. https://doi.org/10.1098/rspb.2002.2239

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods*, *49*(1), 282–293. https://doi.org/10.3758/s13428-015-0697-6

Wiley, R. H. (2013). Signal detection, noise, and the evolution of communication. In H. Brumm (Ed.), *Animal communication and noise* (2nd ed., pp. 7–31). Springer. https://doi.org/10.1007/978-3-642-41494-7_2

Wilkins, M. R., Seddon, N., & Safran, R. J. (2013). Evolutionary divergence in acoustic signals: Causes and consequences. *Trends in Ecology & Evolution*, *28*(3), 156–166. https://doi.org/10.1016/j.tree.2012.10.002

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.