

1 Quantifying similarity in animal vocal sequences: which metric performs best?

2

3 Arik Kershenbaum<sup>1\*</sup>, Ellen C. Garland<sup>2</sup>

4

5 <sup>1</sup>Department of Zoology, University of Cambridge, Cambridge, England

6 <sup>2</sup>School of Biology, University of St. Andrews, St. Andrews, Fife, KY16 9TH, Scotland

7

8 \*Author for correspondence: arik.kershenbaum@gmail.com

9

10 Short title: Measuring sequence similarity

11

12 Word count: 7431

13

14

15 SUMMARY

- 16 1. Many animals communicate using sequences of discrete acoustic elements which can be complex,  
17 vary in their degree of stereotypy, and are potentially open-ended. Variation in sequences can  
18 provide important ecological, behavioural, or evolutionary information about the structure and  
19 connectivity of populations, mechanisms for vocal cultural evolution, and the underlying drivers  
20 responsible for these processes. Various mathematical techniques have been used to form a  
21 realistic approximation of sequence similarity for such tasks.
- 22 2. Here, we use both simulated and empirical datasets from animal vocal sequences (rock hyrax,  
23 *Procavia capensis*; humpback whale, *Megaptera novaeangliae*; bottlenose dolphin, *Tursiops*  
24 *truncatus*; and Carolina chickadee, *Poecile carolinensis*) to test which of eight sequence analysis  
25 metrics are more likely to reconstruct the information encoded in the sequences, and to test the  
26 fidelity of estimation of model parameters, when the sequences are assumed to conform to  
27 particular statistical models.
- 28 3. Results from the simulated data indicated that multiple metrics were equally successful in  
29 reconstructing the information encoded in the sequences of simulated individuals (Markov chains,  
30 n-gram models, repeat distribution, and edit distance), and data generated by different stochastic  
31 processes (entropy rate and n-grams). However, the string edit (Levenshtein) distance performed  
32 consistently and significantly better than all other tested metrics (including entropy, Markov  
33 chains, n-grams, mutual information) for all empirical datasets, despite being less commonly used  
34 in the field of animal acoustic communication.
- 35 4. The Levenshtein distance metric provides a robust analytical approach that should be considered  
36 in the comparison of animal acoustic sequences in preference to other commonly employed  
37 techniques (such as Markov chains, hidden Markov models, or Shannon entropy). The recent  
38 discovery that non-Markovian vocal sequences may be more common in animal communication  
39 than previously thought, provides a rich area for future research that requires non-Markovian  
40 based analysis techniques to investigate animal grammars and potentially the origin of human  
41 language.

42 Keywords: Sequence, animal communication, vocal, edit distance, Markov, stochastic processes

## 43 INTRODUCTION

44 Many animals communicate using sequences of discrete acoustic elements, the best known example  
45 being bird song, which is composed of multiple notes combined in a distinctive order. These  
46 sequences are often complex, non-stereotyped, and potentially open-ended; that is, individuals may  
47 use an almost unlimited repertoire of sequences by making subtle or large variations to the order of  
48 notes (reviewed in Catchpole & Slater 2003). The role of such sequences varies among species. In  
49 some cases, sequences appear to advertise male quality through sequence complexity, e.g., in marsh  
50 warblers, *Acrocephalus palustris* (Darolová *et al.* 2012); zebra finches, *Taeniopygia guttata* (Holveck  
51 *et al.* 2008; Neubauer 1999; Searcy & Andersson 1986); and song sparrows, *Melospiza melodia* (Pfaff  
52 *et al.* 2007). In other cases, researchers have proposed that sequences contain detailed communicative  
53 information such as individual identity, e.g., bottlenose dolphins, *Tursiops truncatus* (Sayigh *et al.*  
54 1999). It is also possible that in some species, acoustic sequences are essentially stochastic with little  
55 significance to their precise composition.

56 Identifying the role of acoustic sequences in a particular species often involves comparing sequences  
57 within and between individuals, as well as within and between populations, so that the nature of the  
58 variation can be quantified and potentially correlated to ecological or behavioural factors. The task of  
59 comparing acoustic sequences presumes an unequivocal and globally relevant measure of sequence  
60 similarity, or difference. However, in practice, no such metric exists. It could be postulated that a  
61 measure of sequence similarity should reflect the proximal processes taking place in the brains of  
62 intended conspecific signal receivers; i.e., the best measure of sequence similarity is the one used by  
63 the animal itself (Kershenbaum *et al.* 2014). Given that such knowledge is essentially hidden in  
64 practice, various mathematical techniques have been used to form a realistic approximation of signal  
65 similarity (Ashby & Perrin 1988; Navarro 2001; Ranjard 2010; Young & Hamer 1994). It is possible  
66 to categorise similarity measures into two distinct approaches. Firstly, it is usually possible to  
67 characterise a sequence by measuring a small number of metrics that are inherent to the sequence  
68 itself; examples of this include length, or entropy (Freeberg & Lucas 2012). Sequences can then be  
69 compared by calculating the sum of square differences between each of these metrics. This is

70 equivalent to representing each sequence as a “feature vector” in some relatively compact feature  
71 space, and measuring the distance between two sequences as the Euclidean distance between their two  
72 feature vectors. While this method is straightforward, there is an assumption that it is possible to  
73 represent every sequence in a compact way, i.e., that some sufficiently large combination of metrics  
74 can “summarise” the properties of a sequence in a biologically meaningful way. However, it is far  
75 from clear that there exists a compact, yet exact, mathematical representation of a sequence, short of  
76 the trivial task of writing down the entire sequence of elements and attempting to measure the  
77 Euclidean distance between the full representations of two sequences, which is unlikely to produce the  
78 desired results. An alternative approach is to use aggregate techniques that measure properties of a  
79 large number of sequences, and summarise the characteristics of a corpus. For example, sequence  
80 transition tables and element frequency histograms have been used in previous studies (Jin &  
81 Kozhevnikov 2011). In these cases, each vector in feature space represents a collection of sequences,  
82 and the Euclidean distance between vectors measures the difference between the sequences from two  
83 sets of vocalisations, rather than between individual sequences. However, it is questionable whether  
84 any of these techniques, individual or aggregate, can represent the nature of the sequences with  
85 adequate fidelity. Since we do not know what cognitive processes an animal uses to interpret such  
86 sequences, we cannot be sure that any particular summary metric accurately reflects the interpretation  
87 of the sequence by the receiving individual. We refer to all of these above metrics as “unary”, as they  
88 are derived from measurements on each string sequence in isolation, even if distances are eventually  
89 calculated on an aggregate of sequences.

90 Secondly, it is possible to measure the difference between a pair of sequences directly (Levenshtein  
91 1966), thereby bypassing the construction of a feature space, and generating a series of pairwise  
92 comparisons between sequences. Analysing the sequence of elements in animal vocalisations can be  
93 considered analogous to analysing the sequence of nucleotides in DNA, and some non-aggregate  
94 techniques have been borrowed from the field of bioinformatics to capture the similarity or difference  
95 between two sequences. This approach provides a direct measure of pairwise differences, in the form  
96 of a distance matrix, but without a Euclidean feature space. We refer to these metrics as “binary”, as

97 they can only be calculated as a pairwise comparison between exactly two sequences. Binary  
98 difference measures are attractive, as they do not rely on the fidelity of a particular unary metric in  
99 representing the properties of a sequence. Rather, binary metrics are an unequivocal measure of the  
100 similarity/difference between two sequences; although it cannot be assumed that this measure of  
101 similarity is the same as that used by the animal itself in distinguishing between sequences. Such  
102 metrics have long been proposed for the analysis of birdsong (Bradley & Bradly 1983; Ranjard *et al.*  
103 2010), but have not been widely adopted. One disadvantage of binary metrics is that a number of  
104 common machine learning algorithms often used for clustering the results of similarity analyses (e.g.,  
105 k-means, neural networks) rely on data presented as a Euclidean feature space, although there are  
106 exceptions, e.g. Ranjard & Ross (2008). To use such clustering techniques, it would be necessary to  
107 derive a series of feature vectors from the binary metric distance matrix. This can be done using  
108 techniques such as multidimensional scaling or principal component analysis to convert a distance  
109 matrix to feature vectors.

110 Here, we compare the performance of eight different methods for analysing animal vocal sequences,  
111 using both aggregate statistical metrics and a direct pairwise distance measure. We use simulated and  
112 empirical sequences to test which approach is more likely to reconstruct the information encoded in  
113 the sequences, and to test the fidelity of estimation of model parameters when the sequences are  
114 assumed to conform to particular statistical models. This direct comparison of a number of commonly  
115 employed analytical algorithms provides a comprehensive evaluation of the utility of these  
116 approaches to real-world data sets, and demonstrates the utility of comparing at least two different  
117 methods when assessing novel algorithms to ensure that results are robust under a range of analytical  
118 approaches.

119

## 120 METHODS

121 We performed two sets of tests (*viz.* artificial and empirical) to evaluate the performance of each  
122 metric. In the first tests, we generated artificial random sequences and used the different similarity

123 metrics to reconstruct the parameters used to generate these sequences, and the stochastic model  
124 types. In the second set of tests, we analysed recordings of animal vocalisations and used both unary  
125 and binary difference metrics to determine contextual information known to exist in these sequences.  
126 We used the signature whistles of the bottlenose dolphin (Kershenbaum, Sayigh & Janik 2013; Sayigh  
127 *et al.* 2007; Sayigh *et al.* 2007), to reconstruct individual identity, and the songs of the rock hyrax,  
128 *Procavia capensis* (Kershenbaum *et al.* 2012), the humpback whale, *Megaptera novaeangliae*  
129 (Garland *et al.* 2012), and the calls of the Carolina chickadee, *Poecile carolinensis* (Freeberg 2012),  
130 to reconstruct geographical dialect. In the case of the hyrax, humpback whale, and chickadee, the calls  
131 consisted of a sequence of discrete acoustic elements. In contrast, bottlenose dolphin whistles are  
132 often produced in isolation (rather than as a sequence of whistles); therefore we analysed the sequence  
133 of frequency modulation components (e.g., up, down, constant) within whistles, taking these  
134 modulation components as the acoustic elements (for more details see Kershenbaum, Sayigh & Janik  
135 2013). In both our analysis of artificial sequences, and empirical animal vocal sequences, we evaluate  
136 a number of similarity metrics, both binary and unary. Humpback whale song recordings are held at  
137 the University of Queensland, Australia, and by Operation Cetaces in Noumea, New Caledonia.  
138 Dolphin whistle recordings are held at Woods Hole Oceanographic Institution (see Data Accessibility  
139 section for contact details). Before providing details of the simulation experiments and empirical data  
140 analysis, we describe each of the metrics used.

141

## 142 Binary metric

### 143 **Levenshtein distance (LD)**

144 The Levenshtein distance (Levenshtein 1966) is a type of *string edit distance metric*, as it provides a  
145 quantitative measurement of the difference between two string sequences regardless of string length.  
146 Specifically, the Levenshtein distance measures the minimum number of point operations (additions,  
147 deletions, and substitutions) needed to convert one string into another (Levenshtein 1966). By  
148 comparing the position of elements within a string and calculating the number of changes that it takes

149 to change one string into the other, this metric relies more on the sequence of elements and less on the  
150 overall structural pattern. It has been used extensively in other fields, e.g., bioinformatics (Likic 2008)  
151 and text search/retrieve (Reis *et al.* 2004), and in a small number of previous studies of animal  
152 sequences (e.g., Garland *et al.* 2012; Garland *et al.* 2013; Kershenbaum *et al.* 2012; Krull *et al.* 2012),  
153 and is related to the better known dynamic time warping algorithm (Buck & Tyack 1993). However,  
154 LD itself remains somewhat unknown in the field of animal acoustic communication. In practice,  
155 string edit distances are often paired with string alignment algorithms or additional standardisations,  
156 particularly when the strings being compared are of different lengths: Figure 1; see Kershenbaum *et*  
157 *al.* (2012) and Garland *et al.* (2012) for additional information on metric calculation. Importantly, the  
158 Levenshtein distance forms the basis of the Needleman-Wunsch string alignment (Likic 2008;  
159 Needleman & Wunsch 1970) that is used extensively in bioinformatics research to compare sections  
160 of DNA. In our implementation of the LD algorithm, we assign an equal cost (of 1) to any correction  
161 operation (addition, deletion, substitution), no cost (0) for a matching element, and no cost for  
162 differences in sequence lengths after optimal alignment.

163 Although other binary metrics exist apart from LD, they are in general unsuitable for the task at hand.  
164 For example, the Hamming distance requires sequences of the same length, and the most frequent  $k$   
165 characters simply provides a count of the most common symbol/element. These therefore provide less  
166 information than the Levenshtein distance metric.

167

## 168 Unary metrics

### 169 **Transition table (TT)**

170 Acoustic sequences have often been modelled as a Markov chain (Berwick *et al.* 2011; Briefer *et al.*  
171 2010; Briefer *et al.* 2010), in which the probability of a particular element occurring depends only on  
172 the preceding element (or sometimes, more than one preceding element). These conditional  
173 probabilities of each element, given the preceding element(s), can be expressed as a transition matrix  
174  $T$ , in which the element  $T_{i,j}$  represents the probability of the element  $j$  occurring after the element  $i$ .



175 For a sequence consisting of  $C$  distinct element types, a  $C \times C$  transition matrix can be estimated from  
 176 empirical data. When comparing two sequences A and B, the similarity between the transition  
 177 matrices  $T_A$  and  $T_B$  is an indication of the similarity between the sequences (Jin & Kozhevnikov  
 178 2011). To calculate a difference metric  $D_{TT} = f(T_A, T_B)$ , we can express each matrix as a  $C^2$   
 179 dimensional feature vector  $V$ , where the elements of the vector are equal to the elements of the  
 180 transition matrix  $T$ , i.e.,  $V = T(\cdot)$ . We then calculate the Euclidean distance between the two vectors  
 181 derived from sequences A and B:

$$D_{TT}(A, B) = \sqrt{\sum (V_A - V_B)^2}$$

182 However, such a metric would not be expected to produce a meaningful measure for sequences  
 183 composed of non-overlapping element types (e.g. ABCABC, and DEFDEF). Therefore we sort  
 184 vectors  $V_A$  and  $V_B$  in order of transition probability before comparison. This allows a comparison of  
 185 transition probability distributions, independent of element type.

186

### 187 **N-gram distribution (NG)**

188 Researchers have previously proposed that an important property of animal sequences is the nature of  
 189 repeating units within the sequence (Cane 1959; Kershenbaum *et al.* 2014; Pruscha & Maurus 1979).

190 A sequence of length  $L$  consists of  $L-n+1$  sub-sequences of length  $n$ . Thus, the five-element sequence  
 191 ABBAC consists of  $5-2+1=4$  two-element sub-sequences: AB, BB, BA, and AC. For a sequence  
 192 consisting of  $C$  distinct element types, there are a total of  $C^n$  distinct  $n$ -element possible sub-  
 193 sequences. The vector of sub-sequence frequencies,  $P(i \in C^n)$  can be considered a feature vector, and  
 194 the distance between two strings calculated in a similar way to that shown above:

$$D_{NG}(A, B) = \sqrt{\sum (P_A - P_B)^2}$$

195 In the following analyses, we chose the n-gram distribution for  $n = 3$ , as this provides a good balance  
196 between coverage and diversity. For a comparison of different length n-grams in analysing birdsong,  
197 see Jin & Kozhevnikov (2011).

198

### 199 **Shannon entropy (SE)**

200 Information theory approaches to analysing animal vocal communication have become popular in  
201 recent years. One metric that is simple to understand and easy to apply is the Shannon entropy  
202 (Shannon *et al.* 1949), and this has been used in a number of studies to measure the complexity of  
203 animal vocal sequences (Da Silva, Piqueira & Vielliard 2000; Doyle *et al.* 2008; McCowan, Hanser &  
204 Doyle 1999; McCowan, Hanser & Doyle 1999; Suzuki, Buck & Tyack 2006). Shannon entropy  
205 measures the unpredictability of a sequence, or the lack of uniformity of a sequence, so that a  
206 completely predictable sequence (e.g., consisting of the same element repeated over and over) would  
207 have an entropy of zero, whereas a completely unpredictable (random) sequence would have an  
208 entropy of one. The equation for Shannon entropy  $H$  is as follows:

$$H = - \sum_{i \in 1 \dots C} P_i \log_C P_i$$

209 where  $P_i$  is the probability of element  $i$ , drawn from a set of the  $C$  elements occurring in the union of  
210 all sequences.

211 Our SE metric compares two sequences by taking the ratio of the Shannon entropies of the sequences  
212 A and B:

$$D_{SE}(A, B) = H_A/H_B \text{ where } H_A < H_B$$

213 Although SE is calculated as a single comparison between single measurements on two sequences (in  
214 contrast to the TT and NG metrics described above, both of which result in multiple measurements on  
215 a single sequence), SE should still be considered a unary metric, because it does not directly measure  
216 the distance between two sequences, but rather the difference in a derived metric from each.

217

218 **Entropy rate (ER)**

219 Entropy rate has been shown to be a useful metric for measuring vocal sequence complexity  
220 (Kershenbaum 2013). Entropy rate is derived from the transition table of a sequence, and can be  
221 thought of as a measure of transition table diversity, i.e., the extent to which different transitions  
222 between notes are of uniform or non-uniform probability. Given a transition table  $T_{i,j}$  as described  
223 above, entropy rate ER is defined as:

$$ER = - \sum_{i \in 1 \dots C} \pi_i \sum_{j \in 1 \dots C} T_{i,j} \log T_{i,j}$$

224 where  $\pi_i$  is the stationary probability of element  $i$ , i.e., the overall probability of  $i$  occurring in the  
225 sequence; see Kershenbaum (2013) for additional information on metric calculation. As with Shannon  
226 entropy, we define a metric  $D_{ER}$  for the difference between sequences A and B:

$$D_{ER}(A, B) = ER_A / ER_B \text{ where } ER_A < ER_B$$

227

228 **Repeat distribution (RD)**

229 The repeat number distribution was used in a recent study to compare the similarity between natural  
230 and synthetic songs of Bengalese finches, *Lonchura striata var. domestica* (Jin & Kozhevnikov  
231 2011). It is an aggregate measure, calculated on a corpus of sequences. For each set of sequences a  
232 histogram is generated showing the probabilities  $P_n$  that any element occurred in isolation ( $n = 1$ ), was  
233 repeated twice ( $n = 2$ ), three times ( $n = 3$ ), and so on. As with the n-gram distribution, we define a  
234 metric that measures the difference between two such histograms, generated from sequences A and B,  
235 where  $P_A$  and  $P_B$  are the feature vectors of sequences A and B, comprising the repeat distributions for  
236 all the elements:

$$D_{RD}(A, B) = \sqrt{\sum (P_A - P_B)^2}$$

237

238 **Mutual information (MI)**

239 Mutual information is an information theory measure that can be applied easily to quantify the  
240 similarity of two sequences. MI combines both measures of the inherent complexity in a sequence  
241 (via Shannon entropy), and the joint entropy of the sequences, which measures the probability that a  
242 particular pair of elements will occur at the same point in two sequences; see Kershenbaum *et al.*  
243 (2012) for additional information on metric calculation. MI is defined as follows:

$$MI = H(A) + H(B) - \sum_i \sum_j p_{i,j} \log p_{i,j}$$

244 where  $H(A)$  is the Shannon entropy of sequence A,  $H(B)$  is the Shannon entropy of sequence B, and  
245  $p_{i,j}$  is the probability that elements  $i$  and  $j$  occur at the same point in sequences A and B. As with  
246 Shannon entropy, we define a metric  $D_{MI}$  for the difference between sequences A and B:

$$D_{MI} = MI_A / MI_B \text{ where } MI_A < MI_B$$

247

248 **Lempel-Ziv (LZ)**

249 The Lempel-Ziv complexity (Lempel & Ziv 1976) is an important algorithm used for data  
250 compression, as it is a measure of the number of distinct patterns in a sequence. As a metric of  
251 sequence complexity and an approximation to Kolmogorov complexity (Evans & Barnett 2002), it is  
252 potentially a useful indicator of the diversity of an animal vocal sequence. Although it has not been  
253 widely used in animal studies, Suzuki, Buck & Tyack (2006) suggested the use of the LZ metric for  
254 the analysis of humpback whale song, and Kershenbaum (2013) showed that the LZ metric  
255 outperformed Shannon entropy (SE) in quantifying realistic length acoustic sequences. LZ complexity  
256 was calculated using the Applied Nonlinear Time Series Analysis library for Matlab (Small 2005).

$$LZ = \frac{c \log L}{L \log K}$$

257 where  $c$  is the number of distinct substrings in a sequence of length  $L$ , and  $K$  is the maximum number  
258 of possible distinct substrings.

## 259 Sequences for analysis

### 260 **Artificial sequences**

261 In the first test, we evaluated the utility of each of the similarity metrics by their ability to identify  
262 correctly the stochastic process model from which artificial sequences were generated. We generated  
263 artificial sequences using three different stochastic processes, often used to model animal vocal  
264 sequences (Kershenbaum *et al.* 2014) : the zero-order Markov process (ZOMP), the first-order  
265 Markov process (FOMP), and the semi-Markov renewal process (RP). The ZOMP is an independent  
266 stochastic process, in which the probability of any particular element occurring at a particular point in  
267 a sequence is determined solely by the prior probability of that element. In the FOMP, element  
268 probabilities are determined by a transition table, where the probability of a particular element  
269 depends on the immediately preceding element. The RP has been shown to be a more realistic model  
270 of animal vocal sequence production (Kershenbaum *et al.* 2014) in which the number of repeated  
271 elements is drawn from a Poisson distribution, rather than being determined by the diagonal of a  
272 transition table. In each case, we examined 10 sequences of 10 elements each, drawn from five  
273 possible elements (A-E). We generated 30 sequences, 10 from each of the stochastic processes,  
274 ZOMP, FOMP, and RP. The ZOMP was modelled by selecting five random prior probabilities, one  
275 for each element type, and renormalising to sum to unity. We then generated the sequences by  
276 selecting elements according to these prior probabilities. The FOMP was modelled by generating a  
277 random 5 x 5 transition table in a similar way to the ZOMP prior probabilities, so that the rows of the  
278 transition matrix summed to unity. A random initial element was chosen for each 10-element  
279 sequence, and the remaining nine elements in each sequence were chosen randomly according to the  
280 probabilities in the transition table. The RP was modelled in a similar way to the FOMP, except that  
281 for each element generated, a random number of repeats were drawn from a Poisson distribution with  
282 mean five (to give 95% confidence of  $\leq 9$  repeats). Having generated 30 sequences of 10 elements, we

283 then calculated a 30 x 30 distance matrix for each of the similarity metrics. We then used an Adaptive  
284 Resonance Theory (ART) artificial neural network to cluster these 30 points into natural groupings,  
285 setting a maximum of 100 possible clusters. ART networks have been used in a number of previous  
286 studies to cluster data derived from animal vocalisations (Deecke & Janik 2006; Janik 1999; Quick &  
287 Janik 2012). We then calculated the normalised mutual information (NMI) as a metric of goodness of  
288 clustering (Zhong & Ghosh 2005), by comparing the composition of the generated clusters  $H(Y)$  with  
289 the true generation process of each  $H(\hat{Y})$ . Thus, NMI indicates the proportion of uncertainty predicted  
290 by the metric. We then repeated this process 100 times using new random transition matrices,  
291 generating 3000 sequences in total.

292 In the second test using artificial sequences, we simulated “individuals” by generating 100 random RP  
293 transition matrices, and from each of them producing a set of 10 sequences of 10 elements each. We  
294 used the RP generation process, rather than a Markovian ZOMP or FOMP, as the RP more reliably  
295 describes many types of animal vocal sequences (Kershenbaum *et al.* 2014). Each sequence generated  
296 from a single transition matrix would be expected to be more similar to other sequences from the  
297 same transition matrix, than sequences generated by a different random transition matrix, therefore we  
298 used a similar clustering approach as in the stochastic process analysis above. We calculated the 100 x  
299 100 distance matrix for each similarity metric, obtained by comparing the sequences from each of the  
300 100 transition matrices, and clustered the results as before, measuring the NMI as an indication of  
301 clustering success.

302 For a final test using artificial sequences, we examined the effect of typical sample sizes (number of  
303 sequences) on each of the similarity metrics. Using the sequences generated in the individual  
304 simulation above, we varied the number of sequences analysed from one to ten, recalculated the  
305 distance matrices and clustering, and measured the NMI.

306

307 **Animal sequences**

308 We tested the performance of the above metrics using empirical sequences of animal vocalisations,  
309 where those sequences are thought to contain information that is known *a priori*. Very few examples  
310 exist where contextual information is objectively known to exist in animal vocal sequences. However,  
311 the signature whistles of bottlenose dolphins have been shown to encode individual identity in the  
312 sequence of up-down frequency shifts, known as a Parsons code (Kershenbaum, Sayigh & Janik  
313 2013). We used a data set consisting of 400 signature whistles, 20 from each of 20 individual  
314 dolphins, recorded during capture-release events; see Sayigh *et al.* (2007) and Kershenbaum, Sayigh  
315 & Janik (2013) for additional details. We converted each whistle into a 9-element Parsons code, with  
316 seven possible element values (“large drop”, “medium drop”, “small drop”, “no change”, “small  
317 rise”, “medium rise”, and “large rise”). We then calculated distance matrices using each of the  
318 similarity metrics described above, and clustered using an ART network. For the calculation of NMI,  
319 we compared the generated clusters to the known clusters of individual identity. As empirical data do  
320 not allow the generation of unlimited data sets as with artificial sequences, we estimated confidence  
321 intervals for each of the empirical data sets by randomly selecting 80% of the calls for clustering and  
322 calculation of NMI, and repeated this process 100 times.

323 We analysed three further empirical data sets for which contextual information in vocal sequences has  
324 been proposed. The first data set used recordings of humpback whales (for details see Garland *et al.*  
325 2012), the second data set used recordings of rock hyraxes (see Kershenbaum *et al.* 2012), and the  
326 third set Carolina chickadees (see Freeberg 2012). Previous studies have shown that in the humpback  
327 whale, rock hyrax, and Carolina chickadee, song syntax varies according to the geographical origin of  
328 the population. For example, not only does chickadee song syntax vary between locations, but there  
329 appear to be different functional use of certain sequences in the different populations (Freeberg 2012).

330 The humpback whale data set consisted of 202 songs composed of 20 different element types  
331 (themes), recorded from 42 individuals. Humpback whale song is a complex, stereotyped, repetitive,  
332 long, male display that has multiple levels of hierarchy in its organisation (Herman & Tavolga 1980;  
333 Payne & Payne 1985; Payne & McVay 1971). A few sounds (units) are arranged in a stereotyped  
334 phrase which is repeated multiple times to make a theme (Payne & McVay 1971). A number of

335 themes, sung in a particular order, are combined to form a song. The order and content of the themes  
336 are highly stereotyped, and all males within a population adhere to the same arrangement and content  
337 of the song at any given time as the display is constantly changing (Frumhoff 1983; Payne, Tyack &  
338 Payne 1983; Payne & Payne 1985). This analysis focused on the theme level in the hierarchical  
339 arrangement of humpback whale song. Each string therefore represented the sequence of themes  
340 (elements) that comprised a song; e.g., theme 1, theme 2, theme 3, theme 4, theme 5; see Garland *et*  
341 *al.* (2012) for further information and example sequences. This level within the hierarchy takes into  
342 account information on the sequence of units and the repetition of phrases at a higher level, but does  
343 not examine these lower levels explicitly. Strings were classified according to their geographical  
344 location: New Caledonia, Vanuatu, or eastern Australia, and this geographical origin was compared to  
345 the clusters generated by the ART network. Humpback whale song is constantly changing, and has  
346 been shown to undergo complete song revolutions in this region (Garland *et al.* 2011; Garland *et al.*  
347 2011; Noad *et al.* 2000). The current analysis incorporates two different song types (lineages) that  
348 contain different themes (vocabulary), and are present in these populations at various points over the  
349 four years of recording. Therefore, each metric must be robust to the underlying transmission  
350 dynamics of this display.

351 The hyrax data consisted of 1130 song sequences composed of five different element types, recorded  
352 from a single individual at each of 18 different locations in Israel. The Carolina chickadee data  
353 consisted of 1184 sequences of calls, recorded from 60 sites in the states of Tennessee and Indiana,  
354 USA. Links to these data sets are available in the supplemental information.

355

## 356 RESULTS

### 357 **Artificial sequences**

358 For sequences generated by different stochastic processes, the entropy rate (ER) metric provided the  
359 best clustering, with a NMI value of  $0.518 \pm 0.005$  (standard error) (Figure 2a), while the binary  
360 Levenshtein distance (LD) metric gave a NMI of  $0.476 \pm 0.006$ . A post-hoc Tukey test following



361 ANOVA showed significant differences between the NMI scores of these two metrics. All other  
362 metrics produced significantly lower NMI values.

363 Results from clustering sequences of simulated "individuals" (sequences generated by stochastic  
364 processes with similar parameters), indicated that NG produced the highest NMI score  $0.751 \pm 0.001$ ,  
365 while the LD, RD, and TT metrics all produced high but slightly lower NMI scores (greater than 0.7;  
366 Figure 2b), with no significant differences among the NMI values of these three metrics.

367 Both the LD and NG metrics that performed well on the above clustering tasks were also robust to  
368 sample size (Figure 3). Most other metrics were also relatively unaffected by sample size. However,  
369 the RD performed poorly at smaller sample sizes ( $\leq 4$ ), and the MI declined with increasing corpus  
370 size ( $> 2$ ).

371

## 372 **Animal sequences**

373 When clustering to reconstruct the individual identity from bottlenose dolphin signature whistles, the  
374 Levenshtein distance (LD) performed significantly better than all other tested metrics, with an NMI of  
375  $0.661 \pm 0.001$  (Figure 4a). The n-gram distribution (NG) also performed well, with an NMI of  $0.63p \pm$   
376  $0.001$ . Clustering of the humpback whale song data to indicate population (geographic) origin,  
377 showed the LD again performed significantly better than all other tested metrics (NMI of  $0.491 \pm$   
378  $0.005$ ; Figure 4b). The NG provided the second best, although significantly poorer, metric (NMI of  
379  $0.367 \pm 0.005$ ). All metrics performed poorly in clustering the geographical origin of hyrax songs;  
380 however, the LD metric was again significantly better than all others tested (NMI  $0.1684 \pm 0.001$ ,  
381 compared to the next best NMI of  $0.130 \pm 0.001$  for TT; Figure 4c). Clustering of the chickadee data  
382 to distinguish between birds recorded in Tennessee and those recorded in Indiana, showed the LD  
383 performed significantly better than all other metrics (NMI of  $0.450 \pm 0.001$ ; Figure 4d), followed by  
384 NG (NMI  $0.369 \pm 0.001$ ).

385

386 DISCUSSION

387 We analysed the performance of eight different techniques from two broad approaches, to investigate  
388 the utility of each approach in the comparison of animal sequences. The unary and binary metrics  
389 performed similarly well in the artificial sequence tests, with the entropy rate (ER) metric slightly  
390 better than the Levenshtein distance binary metric (LD), in distinguishing between data generated by  
391 different stochastic processes, and n-gram (NG) slightly better in distinguishing simulated individuals.  
392 However, the LD metric performed significantly better than all other tested metrics when presented  
393 with empirical animal sequences. This result emphasises that caution should be used when using  
394 artificially generated sequences based on simple stochastic models to simulate animal vocal  
395 sequences. Recent work has shown that assumptions of simple models for animal vocal production are  
396 likely to be inaccurate (Kershenbaum *et al.* 2014), and similar conclusions have been indicated for  
397 cetacean song (Miksis-Olds *et al.* 2008). The difference between metric performance on artificial and  
398 on empirical data is striking. Little is known of the cognitive mechanisms by which animals encode  
399 and decode information in vocalisations (Thornton, Clayton & Grodzinski 2012); researchers must  
400 rely on isolated examples where information content is known *a priori* to draw conclusions about  
401 which analytical techniques are best suited for vocal sequence data. Our results clearly show that the  
402 LD metric outperforms other metrics on empirical data, despite performing less effectively on  
403 simulated data. This indicates that the sequential order of the sequences varied across  
404 location/individual while the level of complexity is similar. The Levenshtein distance was the metric  
405 of choice for clustering dolphin signature whistles into individuals, humpback whale song into  
406 populations, hyrax songs into geographical region, and chickadee calls into state of origin. Analysis of  
407 the sensitivity of the different metrics to sample size showed that most of the metrics that performed  
408 well across the data sets (LD, NG, LZ), were also robust to sample size.

409 Results from the current paper in combination with previous work (Eriksen *et al.* 2005; Garland *et al.*  
410 2012; Garland *et al.* 2013; Helweg *et al.* 1998; Tougaard & Eriksen 2006), highlight the success of  
411 the Levenshtein distance (LD) metric in the analysis of sequence content and comparison of  
412 humpback whale song. A large body of work has previously shown that song differences among

413 humpback whale populations can indicate geographic origin of a singer (e.g., Garland *et al.* 2015;  
414 Helweg *et al.* 1998; Payne & Guinee 1983). Despite dynamic song transmission in the South Pacific  
415 region, fine-scale song differences allow the identification of population origin (Garland *et al.* 2011;  
416 Garland *et al.* 2012; Garland *et al.* 2013; Garland *et al.* 2015). The current paper examined the theme  
417 sequences (i.e., a set of phrases under a single label) as part of the largest analysis to date of sequence  
418 comparison algorithms for humpback whale song (Garland *et al.* 2013), which indicated the LD out  
419 performed all other tested metrics. We suggest when *comparing* song sequences, the LD metric  
420 should be employed preferentially, while if the complexity or information content of each song is the  
421 focus of study, the researcher should employ other techniques such as entropy.

422 Previous studies of sequence comparison in hyrax song (Kershenbaum *et al.* 2012) have shown  
423 geographical variation in sequence structure using the LD metric, as these findings were supported by  
424 application of an unrelated (unary) metric, mutual information (MI). In the current study, MI  
425 performed very poorly on both simulated and empirical data, although MI performance was somewhat  
426 better on the hyrax data than on the other data sets. This implies that the aspect of the sequences that  
427 is measured by MI does not vary in correlation with geographic location or individual. While not all  
428 studies can compare large numbers of analytical algorithms, this emphasises the utility of comparing  
429 at least two different techniques when assessing novel algorithms, to ensure that results are robust  
430 under a range of analytical approaches.

431 Despite all tested metrics performing poorly in the assessment of geographic origin in hyrax song, the  
432 LD metric was significantly better than all others. In previous work, (Kershenbaum *et al.* 2012)  
433 measured the correlation between sequence similarity and the distance between populations, rather  
434 than classification success, and the latter suggests that distinct dialects are not present in the hyrax.  
435 Rather, small but significant differences are present between all pairs of populations, depending on  
436 geographic isolation. In contrast, humpback whales, chickadees, and bottlenose dolphins show strong  
437 discrimination between in-group and out-group sequences, indicating that the differences between the  
438 vocal sequences of different individuals or populations are much more marked. This may indicate an  
439 adaptive role to distinctive vocalisations in dolphins and whales, such as individual identification

440 (Janik & Slater 1998; Janik, Sayigh & Wells 2006; Quick & Janik 2012), while in chickadees  
441 adaptive, developmental, and phylogenetic explanations for regional dialects have been suggested  
442 (Freeberg 2012). Humpback whale song is hypothesised to contain information about the reproductive  
443 fitness and population origin of the signaller (Helweg *et al.* 1992; Helweg *et al.* 1992; Payne &  
444 Guinee 1983). Hyrax song complexity is not thought to contain contextual information beyond male  
445 fitness (Demartsev *et al.* 2014; Koren & Geffen 2009), although this assumption is currently untested.  
446 In contrast, dolphin signature whistles are known to be individually distinctive whistles that can be  
447 identified by the unique pattern of frequency modulations (Janik, Sayigh & Wells 2006). The  
448 characterisation of signature whistles based on a 7-element Parsons code in a previous study  
449 (Kershenbaum, Sayigh & Janik 2013) allows individual identification of the whistler. The LD  
450 significantly outperformed all other models in clustering to reconstruct not only the individual identity  
451 from signature whistles, but the geographic origin for humpback whale song, chickadee calls, and  
452 hyrax song, highlighting the importance of evaluating different metrics with *a priori* information.

453 One likely explanation for the higher performance of the LD metric is that it alone among the metrics  
454 analysed uses a direct comparison of the vocal sequences between samples, thereby using more  
455 information about the sequences than the other metrics. The LD metric by design can solely be  
456 employed to *compare* two strings and it excels at this task; it does not provide an understanding of the  
457 information content within each string, or the sequence structure. By necessity this means that LD  
458 also compares the vocabularies of a pair of sequences, and therefore two sequences that are based on  
459 the same set of sequence elements are likely to have a lower LD value than two sequences that are  
460 composed of different elements, but have similar sequence structure. Regional differences in the  
461 vocabulary (e.g., humpback song themes) provide important information on the connectivity of  
462 populations at a broad-scale despite an overall similarity in song structure (hierarchical arrangement).  
463 To establish the influence of overlapping vocabulary is beyond the scope of this paper (although two  
464 of the three humpback populations switched between two vocabularies – song types – over the course  
465 of this study), but we present as supplemental information (Figure S1) the element distributions of the  
466 different data sets, which in most cases were quite consistent.

467 Sample sizes can be constrained in the study of wild animals and particularly in marine mammal  
468 studies. Samples may be collected infrequently and with a patchy distribution due to the challenging  
469 conditions presented in collecting such data. Understanding how a metric reacts to a small sample size  
470 is invaluable in metric choice. The robust nature of the LD and NG to smaller sample sizes and their  
471 high performance in the comparison task makes them appealing for analysis. The data presented here  
472 indicated that LD and NG performed well with a sample size of three or less, while TT and RD should  
473 not be considered as a metric for analysis until a sample size of four or more is available.

474 Here, we have presented a robust understanding of which metric should be preferentially employed in  
475 studies involving the comparison of individual- or group-specific vocalisations, such as signature  
476 whistles. The success in identifying individual/geographic variations in vocal sequences has  
477 implications for assessing population structure, song transmission, and dialect similarity, particularly  
478 for populations where rapid song changes occur. For example, the analysis of humpback whale song  
479 presented here was able to identify population origin despite rapid song dynamics (Garland *et al.*  
480 2011; Garland *et al.* 2012; Garland *et al.* 2013). We suggest that the LD can be applied to any level  
481 within a complex display, but suggest future studies strive for the lowest level sequence within the  
482 hierarchy (i.e., sequence of units or phrases), to increase the amount of information directly compared  
483 and thus encapsulated by the sequence.

484 The LD method provides a metric to compare sequence content and organisation (and thus songs)  
485 within and among multiple individuals, populations, years, and locations. In particular, transmission  
486 of humpback whale song is largely cultural, and the level and rate of change remains unparalleled in  
487 any other non-human animal as complete population-wide changes are replicated in multiple  
488 populations at a vast geographic scale (Garland *et al.* 2011). Thus, fundamental questions in animal  
489 culture, vocal learning, and cultural evolution can be explored using humpback whale song as a  
490 model, and with the help of the LD metric. Further, the evolution of complex vocal labels (i.e.,  
491 signature whistles) and the underlying cognitive abilities required for such evolution, are extremely  
492 important in understanding the evolution of vocal complexity (Janik 2014). Robust metrics that  
493 capture the information encoded in the sequences with the highest fidelity are thus required to address

494 these far-reaching evolutionary questions. We suggest the LD should be utilised in such comparison  
495 studies in preference to Markov and information theory based models.

496

## 497 Conclusions

498 The Levenshtein distance (LD; binary metric) significantly outperformed all other tested metrics in  
499 our comparative analysis of animal acoustic sequences. It provides a direct measure of pairwise  
500 differences among sequences, instead of a comparison of aggregate similarity. N-grams (Markov  
501 chains) were the second most successful metric; the underlying issue that the tested species'  
502 vocalisations may be governed by non-Markovian dynamics and the consistent success of the LD  
503 metric, suggests n-grams should always be a second choice. Given the inherent interest in the origins  
504 of human language and the evolution of signalling complexity, robust and reliable metrics that can  
505 capture the content and arrangement of the signal are essential to address these fundamental questions  
506 in animal communication and cultural evolution.

507

## 508 Acknowledgements

509 We thank Melinda Rekdahl, Todd Freeberg and his graduate students, Amiyaal Ilany, Elizabeth  
510 Hobson, and Jessica Crance for providing comments of on a previous version of this manuscript. We  
511 thank Mike Noad, Melinda Rekdahl, and Claire Garrigue for assistance with humpback whale song  
512 collection and initial categorisation of the song, Vincent Janik and Laela Sayigh for assistance with  
513 signature whistle collection, Todd Freeberg with chickadee recordings, and Eli Geffen and Amiyaal  
514 Ilany for assistance with hyrax song collection and analysis. E.C.G is supported by a Newton  
515 International Fellowship. Part of this work was conducted while E.C.G. was supported by a National  
516 Research Council (National Academy of Sciences) Postdoctoral Fellowship at the National Marine  
517 Mammal Laboratory, AFSC, NMFS, NOAA. The findings and conclusions in this paper are those of  
518 the authors and do not necessarily represent the views of the National Marine Fisheries Service. We

519 would also like to thank Randall Wells and the Sarasota Dolphin Research Program for the  
520 opportunity to record the Sarasota dolphins, where data were collected under a series of National  
521 Marine Fisheries Service Scientific Research Permits issued to Randall Wells. A.K. is supported by  
522 the Herchel Smith Postdoctoral Fellowship Fund. Part of this work was conducted while A.K. was a  
523 Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute  
524 sponsored by the National Science Foundation through NSF Award #DBI-1300426, with additional  
525 support from The University of Tennessee, Knoxville.

526

#### 527 DATA ACCESSIBILITY

528 Hyrax and chickadee data:

529 [http://rspb.royalsocietypublishing.org/highwire/filestream/47311/field\\_highwire\\_adjunct\\_files/1/rspb](http://rspb.royalsocietypublishing.org/highwire/filestream/47311/field_highwire_adjunct_files/1/rspb)  
530 20141370supp2.xls

531 Authors do not own all of the data and have not been granted permission to archive it. Access is  
532 available as follows:

533 Dolphin data: Dolphin whistle data are held by Dr Sayigh (Woods Hole Oceanographic Institution;  
534 [lsayigh@whoi.edu](mailto:lsayigh@whoi.edu)) and Dr Janik (University of St Andrews; [vj@st-andrews.ac.uk](mailto:vj@st-andrews.ac.uk)). Please contact the  
535 PI's directly for access to their recordings.

536 Humpback whale data: Humpback whale song data are held by Dr. Noad (University of Queensland,  
537 Australia; [mnoad@uq.edu.au](mailto:mnoad@uq.edu.au)) and Dr. Garrigue (Operation Cetaces, New Caledonia;  
538 [op.cetaces@lagoon.nc](mailto:op.cetaces@lagoon.nc)). Please contact the PI's directly for access to their recordings.

539

540 REFERENCES

541

542

543 Ashby, F.G. & Perrin, N.A. (1988). Toward a unified theory of similarity and recognition.

544 *Psychological review*, **95**, 124.

545 Berwick, R.C., Okanoya, K., Beckers, G.J.L. & Bolhuis, J.J. (2011). Songs to syntax: the linguistics

546 of birdsong. *Trends in Cognitive Sciences*, **15**, 113-121.

547 Bradley, D.W. & Bradley, R. (1983). Application of sequence comparison to the study of bird songs.

548 *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison/edited*

549 *by David Sankoff and Joseph B.Krustal,*.

550 Briefer, E., Osiejuk, T.S., Rybak, F. & Aubin, T. (2010). Are bird song complexity and song sharing

551 shaped by habitat structure? An information theory and statistical approach. *Journal of Theoretical*

552 *Biology*, **262**, 151-164.

553 Buck, J.R. & Tyack, P.L. (1993). A quantitative measure of similarity for *Tursiops truncatus*

554 signature whistles. *The Journal of the Acoustical Society of America*, **94**, 2497-2506.

555 Cane, V.R. (1959). Behaviour sequences as semi-Markov chains. *Journal of the Royal Statistical*

556 *Society.Series B (Methodological)*, **21**, 36-58.

557 Catchpole, C.K. & Slater, P.J.B. (2003). *Bird song: biological themes and variations*. Cambridge

558 Univ Press, Cambridge.

559 Da Silva, M.L., Piqueira, J.R.C. & Vielliard, J.M.E. (2000). Using Shannon entropy on measuring the

560 individual variability in the rufous-bellied thrush *Turdus rufiventris* vocal communication. *Journal of*

561 *Theoretical Biology*, **207**, 57-64.



562 Darolová, A., Krištofík, J., Hoi, H. & Wink, M. (2012). Song complexity in male marsh warblers:  
563 does it reflect male quality? *Journal of Ornithology*, **153**, 431-439.

564 Deecke, V.B. & Janik, V.M. (2006). Automated categorization of bioacoustic signals: avoiding  
565 perceptual pitfalls. *The Journal of the Acoustical Society of America*, **119**, 645-653.

566 Demartsev, V., Kershenbaum, A., Ilany, A., Barocas, A., Ziv, E.B., Koren, L. *et al.* (2014). Male  
567 hyraxes increase song complexity and duration in the presence of alert individuals. *Behavioral*  
568 *Ecology*.

569 Doyle, L.R., McCowan, B., Hanser, S.F., Chyba, C., Bucci, T. & Blue, J.E. (2008). Applicability of  
570 information theory to the quantification of responses to anthropogenic noise by southeast Alaskan  
571 humpback whales. *Entropy*, **10**, 33-46.

572 Eriksen, N., Miller, L.A., Tougaard, J. & Helweg, D.A. (2005). Cultural change in the songs of  
573 humpback whales (*Megaptera novaeangliae*) from Tonga. *Behaviour*, **142**, 305-328.

574 Evans, S.C. & Barnett, B. (2002). *Network security through conservation of complexity*. Proceedings  
575 of IEEE MILCOM 2002.

576 Freeberg, T.M. (2012). Geographic variation in note composition and use of chick-a-dee calls of  
577 Carolina chickadees (*Poecile carolinensis*). *Ethology*, **118**, 555-565.

578 Freeberg, T.M. & Lucas, J.R. (2012). Information theoretical approaches to chick-a-dee calls of  
579 Carolina chickadees (*Poecile carolinensis*). *Journal of Comparative Psychology*, **126**, 68-81.

580 Frumhoff, P. (1983). Aberrant songs of humpback whales (*Megaptera novaeangliae*): Clues to the  
581 structure of humpback songs. *Communication and Behavior of Whales* (ed R. Payne), pp. 81-127.  
582 Westview Press, Boulder, Colorado.

583 Garland, E.C., Goldizen, A.W., Lilley, M.S., Rekdahl, M.L., Garrigue, C., Constantine, R. *et al.*  
584 (2015). Population structure of humpback whales in the western and central South Pacific Ocean as  
585 determined by vocal exchange among populations. *Conservation Biology*,.

586 Garland, E.C., Goldizen, A.W., Rekdahl, M.L., Constantine, R., Garrigue, C., Hauser, N.D. *et al.*  
587 (2011). Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale.  
588 *Current Biology*, **21**, 687-691.

589 Garland, E.C., Lilley, M.S., Goldizen, A.W., Rekdahl, M.L., Garrigue, C. & Noad, M.J. (2012).  
590 Improved versions of the Levenshtein distance method for comparing sequence information in  
591 animals' vocalisations: tests using humpback whale song. *Behaviour*, **149**, 1413-1441.

592 Garland, E.C., Noad, M.J., Goldizen, A.W., Lilley, M.S., Rekdahl, M.L., Garrigue, C. *et al.* (2013).  
593 Quantifying humpback whale song sequences to understand the dynamics of song exchange at the  
594 ocean basin scale. *The Journal of the Acoustical Society of America*, **133**, 560-569.

595 Helweg, D.A., Cato, D.H., Jenkins, P.F., Garrigue, C. & McCauley, R.D. (1998). Geographic  
596 variation in South Pacific humpback whale songs. *Behaviour*, **135**, 1-27.

597 Helweg, D.A., Frankel, A.S., Mobley Jr, J.R. & Herman, L.M. (1992). Humpback whale song: Our  
598 current understanding. *Marine Mammal Sensory Systems*, pp. 459-483. Springer.

599 Herman, L.M. & Tavolga, W.N. (1980). The communication systems of cetaceans. *Cetacean*  
600 *behavior: Mechanisms and functions*,, 149-209.

601 Holveck, M., de Castro, Ana Catarina Vieira, Lachlan, R.F., ten Cate, C. & Riebel, K. (2008).  
602 Accuracy of song syntax learning and singing consistency signal early condition in zebra finches.  
603 *Behavioral Ecology*, **19**, 1267-1281.

604 Janik, V.M. (1999). Pitfalls in the categorization of behaviour: a comparison of dolphin whistle  
605 classification methods. *Animal Behaviour*, **57**, 133-143.

606 Janik, V.M., Sayigh, L. & Wells, R. (2006). Signature whistle shape conveys identity information to  
607 bottlenose dolphins. *Proceedings of the National Academy of Sciences*, **103**, 8293-8297.

608 Janik, V.M. & Slater, P.J.B. (1998). Context-specific use suggests that bottlenose dolphin signature  
609 whistles are cohesion calls. *Animal Behaviour*, **56**, 829-838.

610 Janik, V.M. (2014). Cetacean vocal learning and communication. *Current Opinion in Neurobiology*,  
611 **28**, 60-65.

612 Jin, D.Z. & Kozhevnikov, A.A. (2011). A compact statistical model of the song syntax in Bengalese  
613 finch. *PLoS Computational Biology*, **7**, e1001108.

614 Kershenbaum, A. (2013). Entropy rate as a measure of animal vocal complexity. *Bioacoustics*, **23**,  
615 195-208.

616 Kershenbaum, A., Ilany, A., Blaustein, L. & Geffen, E. (2012). Syntactic structure and geographical  
617 dialects in the songs of male rock hyraxes. *Proceedings of the Royal Society B: Biological Sciences*,  
618 **279**, 2974-2981.

619 Kershenbaum, A., Sayigh, L.S. & Janik, V.M. (2013). The encoding of individual identity in dolphin  
620 signature whistles: how much information is needed? *PLoS One*, **8**, e77671.

621 Kershenbaum, A., Blumstein, D.T., Roch, M.A., Akçay, Ç, Backus, G., Bee, M.A. *et al.* (2014).  
622 Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*,.

623 Kershenbaum, A., Bowles, A.E., Freeberg, T.M., Jin, D.Z., Lameira, A.R. & Bohn, K. (2014). Animal  
624 vocal sequences: not the Markov chains we thought they were. *Proceedings of the Royal Society B:*  
625 *Biological Sciences*, **281**, 20141370.

626 Koren, L. & Geffen, E. (2009). Complex call in male rock hyrax (*Procavia capensis*): a multi-  
627 information distributing channel. *Behavioral Ecology and Sociobiology*, **63**, 581-590.

628 Krull, C., Ranjard, L., Landers, T., Ismar, S., Matthews, J. & Hauber, M. (2012). Analyses of sex and  
629 individual differences in vocalizations of Australasian gannets using a dynamic time warping  
630 algorithm. *The Journal of the Acoustical Society of America*, **132**, 1189.

631 Lempel, A. & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on*  
632 *Information Theory*, **22**, 75-81.

633 Levenshtein, V.I. (1966). *Binary codes capable of correcting deletions, insertions and reversals.*  
634 *Soviet Physics Doklady*, **10**, 707-710

635 Likic, V. (2008). *The Needleman-Wunsch algorithm for sequence alignment*,  
636 <http://www.ludwig.edu.au/course/lectures2005/likic.pdf> edn. Lecture given at the 7th Melbourne  
637 Bioinformatics Course, University of Melbourne.

638 McCowan, B., Hanser, S.F. & Doyle, L.R. (1999). Quantitative tools for comparing animal  
639 communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal*  
640 *Behaviour*, **57**, 409-419.

641 Miksis-Olds, J.L., Buck, J.R., Noad, M.J., Cato, D.H. & Stokes, M.D. (2008). Information theory  
642 analysis of Australian humpback whale song. *The Journal of the Acoustical Society of America*, **124**,  
643 2385-2393.

644 Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*,  
645 **33**, 31-88.

646 Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities  
647 in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443-453.

648 Neubauer, R.L. (1999). Super-normal length song preferences of female zebra finches (*Taeniopygia*  
649 *guttata*) and a theory of the evolution of bird song. *Evolutionary Ecology*, **13**, 365-380.

650 Noad, M.J., Cato, D.H., Bryden, M., Jenner, M. & Jenner, K.C.S. (2000). Cultural revolution in whale  
651 songs. *Nature*, **408**, 537-537.

652 Payne, K. & Payne, R. (1985). Large scale changes over 19 years in songs of humpback whales in  
653 Bermuda. *Zeitschrift für Tierpsychologie*, **68**, 89-114.

654 Payne, K., Tyack, P. & Payne, R. (1983). Progressive changes in the songs of humpback whales  
655 (*Megaptera novaeangliae*): a detailed analysis of two seasons in Hawaii. *Communication and*  
656 *Behavior of Whales*, (ed. R. Payne), Westview Press, Boulder, pp 9-57.

657 Payne, R.S. & McVay, S. (1971). Songs of humpback whales. *Science*, **173**, 585-597.

658 Payne, R. & Guinee, L.N. (1983). Humpback whale (*Megaptera novaeangliae*) songs as an indicator  
659 of “stocks”. *Communication and Behavior of Whales*, (ed. R. Payne), Westview Press, Boulder, pp,  
660 333-358.

661 Pfaff, J.A., Zann, L., MacDougall-Shackleton, S.A. & MacDougall-Shackleton, E.A. (2007). Song  
662 repertoire size varies with HVC volume and is indicative of male quality in song sparrows (*Melospiza*  
663 *melodia*). *Proceedings of the Royal Society B: Biological Sciences*, **274**, 2035-2040.

664 Pruscha, H. & Maurus, M. (1979). Analysis of the temporal structure of primate communication.  
665 *Behaviour*, **69**, 118-134.

666 Quick, N.J. & Janik, V.M. (2012). Bottlenose dolphins exchange signature whistles when meeting at  
667 sea. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 2539-2545.

668 Ranjard, L. (2010). *Computational biology of bird song evolution*. PhD, University of Auckland.

669 Ranjard, L., Anderson, M.G., Rayner, M.J., Payne, R.B., McLean, I., Briskie, J.V. *et al.* (2010).  
670 Bioacoustic distances between the begging calls of brood parasites and their host species: a  
671 comparison of metrics and techniques. *Behavioral Ecology and Sociobiology*, **64**, 1915-1926.

672 Ranjard, L. & Ross, H.A. (2008). Unsupervised bird song syllable classification using evolving neural  
673 networks. *The Journal of the Acoustical Society of America*, **123**, 4358-4368.

674 Reis, D.d.C., Golgher, P.B., Silva, A. & Laender, A. (2004). *Automatic web news extraction using*  
675 *tree edit distance*. ACM.

676 Sayigh, L.S., Esch, H.C., Wells, R.S. & Janik, V.M. (2007). Facts about signature whistles of  
677 bottlenose dolphins, *Tursiops truncatus*. *Animal Behaviour*, **74**, 1631-1642.

678 Sayigh, L.S., Tyack, P.L., Wells, R.S., Solow, A.R., Scott, M.D. & Irvine, A.B. (1999). Individual  
679 recognition in wild bottlenose dolphins: a field test using playback experiments. *Animal Behaviour*,  
680 **57**, 41-50.

681 Searcy, W.A. & Andersson, M. (1986). Sexual selection and the evolution of song. *Annual Review of*  
682 *Ecology and Systematics*, **17**, 507-533.

683 Shannon, C.E., Weaver, W., Blahut, R.E. & Hajek, B. (1949). *The Mathematical Theory of*  
684 *Communication*. University of Illinois Press, Urbana.

685 Small, M. (2005). *Applied nonlinear time series analysis: applications in physics, physiology and*  
686 *finance*. World Scientific Publishing Company Incorporated, Singapore.

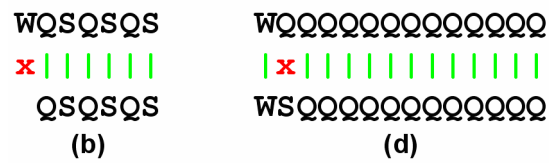
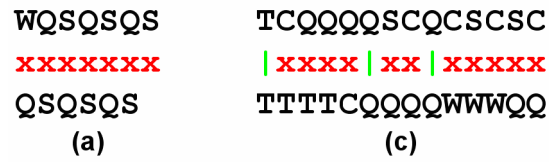
687 Suzuki, R., Buck, J.R. & Tyack, P.L. (2006). Information entropy of humpback whale songs. *The*  
688 *Journal of the Acoustical Society of America*, **119**, 1849-1866.

689 Thornton, A., Clayton, N.S. & Grodzinski, U. (2012). Animal minds: from computation to evolution.  
690 *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 2670-2676.

691 Tougaard, J. & Eriksen, N. (2006). Analysing differences among animal songs quantitatively by  
692 means of the Levenshtein distance measure. *Behaviour*, **143**, 239-252.

- 693 Young, F. & Hamer, R. (1994). Theory and applications of multidimensional scaling. *Hillsdale, NJ:*  
694 *Eribaum Associates,*.
- 695 Zhong, S. & Ghosh, J. (2005). Generative model-based document clustering: a comparative study.  
696 *Knowledge and Information Systems*, **8**, 374-384.
- 697

699



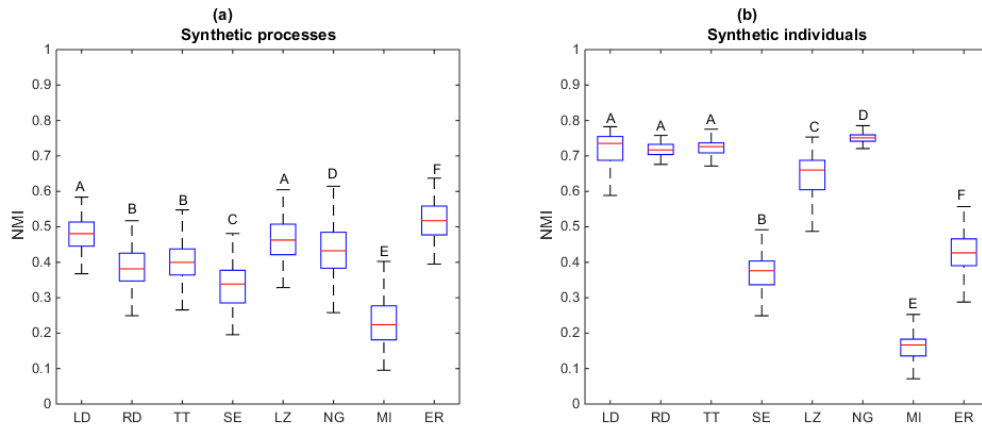
700

701 Figure 1. Examples of string alignment and edit distance. (a) Two unaligned strings with a LD of 7.  
 702 (b) After aligning the strings to minimise the difference, LD = 1. (c) Two hyrax bouts which are  
 703 highly different, LD = 11. (d) Two bouts which are very similar, LD = 1. Reproduced from  
 704 (Kershenbaum *et al.* 2012).

705



706

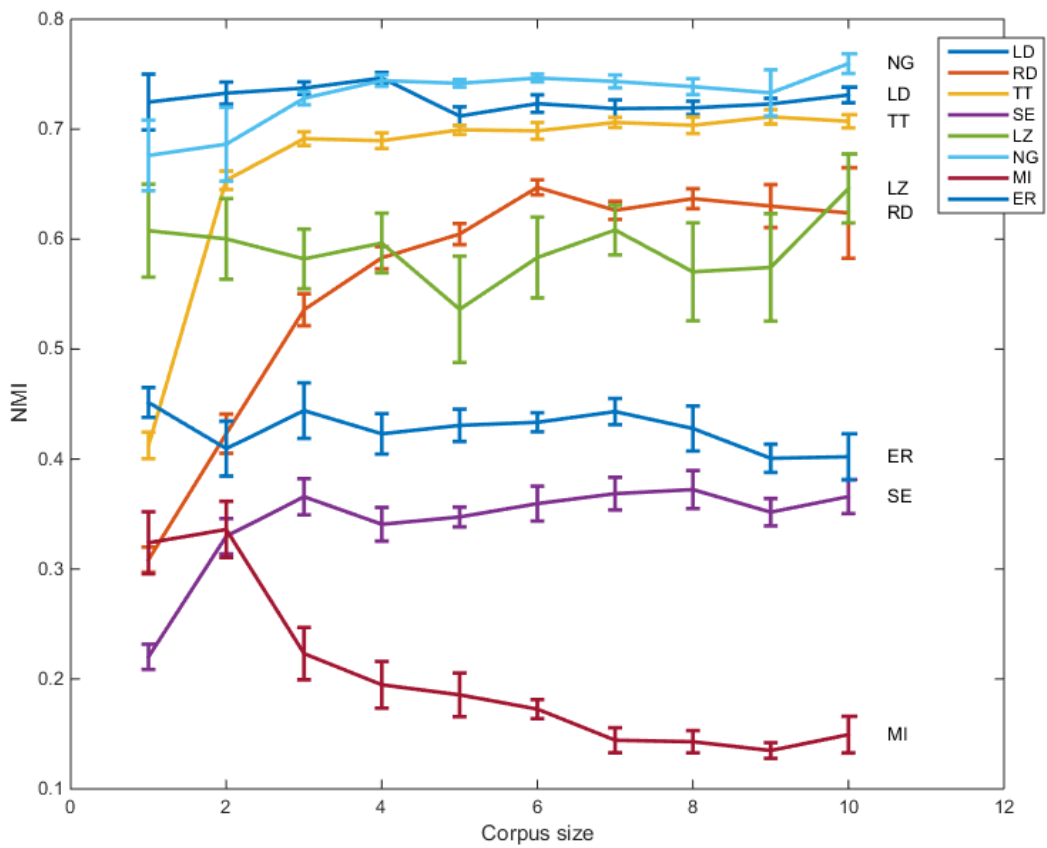


707

708 Figure 2. Results of the normalised mutual information (NMI) scores for each metric using a)  
709 synthetic processes, and b) synthetic individuals. Metric labels: Levenshtein distance (LD), Repeat  
710 distribution (RD), Transition table (TT), Shannon entropy (SE), Lempel-Ziv (LZ), N-gram (NG),  
711 Mutual information (MI), and entropy rate (ER). A-F indicate post-hoc Tukey groupings.

712

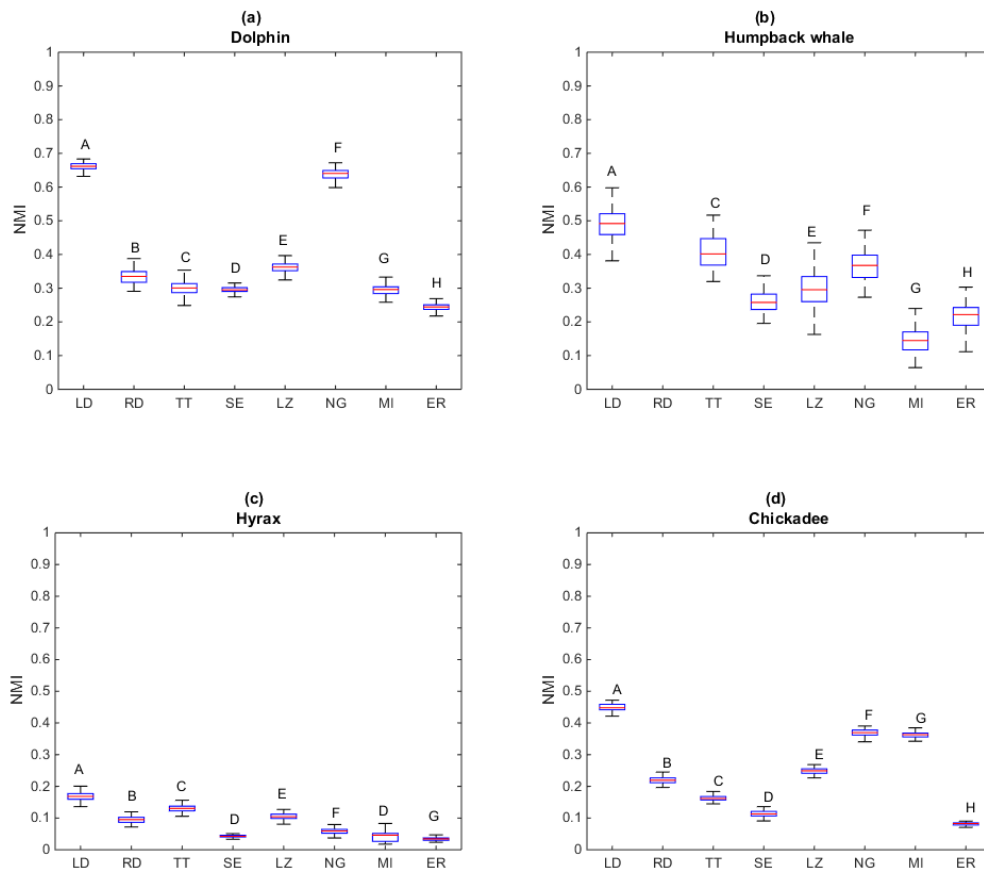
713



714

715 Figure 3. Results of the effect of sample (corpus) size on the NMI scores ( $\pm$  standard error) for each  
716 similarity metric. Metric labels are the same as Figure 2.

717



720 Figure 4. Results of the NMI (normalised mutual information) scores for each metric using a)  
 721 bottlenose dolphin signature whistles, b) humpback whale songs, c) rock hyrax songs, and d) Carolina  
 722 chickadee calls. Metric labels are the same as Figure 2. A-F indicate post-hoc Tukey groupings.