A novel statistical method for behaviour sequence analysis and its application to birdsong

Sarah J. Alger[a,b,c], **Bret R. Larget**[c,d], and **Lauren V. Riters**[b,*]

[a] Department of Biology, University of Wisconsin-Stevens Point, WI, U.S.A.

[b] Department of Zoology, University of Wisconsin-Madison, WI, U.S.A.

[c] Department of Statistics, University of Wisconsin-Madison, WI, U.S.A.

[d] Department of Botany, University of Wisconsin-Madison, WI, U.S.A.

Abstract

Complex vocal signals, such as birdsong, contain acoustic elements that differ in both order and duration. These elements may convey socially relevant meaning, both independently and through their interactions, yet statistical methods that combine order and duration data to extract meaning have not, to our knowledge, been fully developed. Here we design novel semi-Markov methods, Bayesian estimation and classification trees to extract order and duration information from behavioural sequences and apply these methods to songs produced by male European starlings, *Sturnus vulgaris*, in two social contexts in which the function of song differs: a spring (breeding) and autumn (nonbreeding) context. Additionally, previous data indicate that damage to the medial preoptic nucleus (POM), a brain area known to regulate male sexually motivated behaviour, affects structural aspects of starling song such that males in a sexually relevant context (i.e. spring) sing shorter songs than appropriate for this context. We further test the utility of our statistical approach by comparing attributes of song structure in POM-lesioned males to song produced by control spring and autumn males. Spring and autumn songs were statistically separable based on the duration and order of phrase types. Males produced more structurally complex aspects of song in spring than in autumn. Spring song was also longer and more stereotyped than autumn song, both attributes used by females to select mates. Songs produced by POM-lesioned males in some cases fell between measures of spring and autumn songs but differed most from songs produced by autumn males. Overall, these statistical methods can effectively extract biologically meaningful information contained in many behavioural sequences given sufficient sample sizes and replication numbers.

* Correspondence: L. V. Riters, Department of Zoology, 428 Birge Hall, 430 Lincoln Drive, University of Wisconsin-Madison, Madison, WI 53706, U.S.A. LVRiters@wisc.edu (L. V. Riters).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Keywords

In many vertebrates, vocal communication is central to the coordination of social behaviour. This has been particularly well studied in songbirds, which adjust vocal sequences contained in songs to convey meaning about social status and motivational state (Catchpole & Slater, 2008). Vocal elements can differ in the duration (i.e. the length) of time that each element is produced in such a way that the durations of elements are not necessarily independent of their order. The importance of the inextricable relationship between duration and order in sequenced behaviours has not been well studied in part because statistical tools for doing so are not readily available. The goal of the present study was to develop statistical tools to fill this gap and to apply them to basic questions about vocal communication in songbirds.

Some seasonally breeding birds, including European starlings, *Sturnus vulgaris*, sing throughout the year, but the function of song and aspects of song structure change seasonally. Starlings and other songbirds show seasonal changes in broad measures of song structure such that songs produced during the breeding season are often longer (Eens, Pinxten, & Verheyen, 1991; Lampe & Espmark, 1987; Riters et al., 2000; Van Hout, Eens, Balthazart, & Pinxten, 2009), contain a larger repertoire of song element types (Eens et al., 1991) and are more stereotyped by some measures (Nottebohm, Nottebohm, & Crane, 1986; Smith, Brenowitz, Beecher, & Wingfield, 1997) compared to song produced outside of the breeding season. Structural features of male song appear vital to reproductive success. Females of several species prefer as mates males that produce relatively long songs (Bensch & Hasselquist, 1992; Eens et al., 1991; Gentner & Hulse, 2000; Nolan & Hill, 2004) with larger repertoires (Baker, Bjerke, Lampe, & Espmark, 1986; Eens et al., 1991; Searcy & Marler, 1981) and a high degree of stereotypy (Woolley & Doupe, 2008).

Birdsong consists of species-specific arrangements of acoustic elements. The organization of these elements can vary widely across species, but in general the simplest elements of song are referred to as 'notes'. A series of notes that occurs together regularly is referred to as a 'syllable', and a repeated collection of syllables is referred to as a song 'phrase' or 'motif'. Collections of similar phrases or motifs are referred to as song 'types', which when separated by a gap of silence are referred to as 'bouts' (reviewed in Brenowitz, Margoliash, & Nordeen, 1997). In some species, song bouts can be divided into relatively distinct categories containing similar phrase types, as described in detail below for starlings (Eens, 1997; Hausberger, 1997; Mountjoy & Lemon, 1995). Birds may vary the sequential ordering of phrase categories and the amount of time spent producing distinct phrase categories. This suggests the possibility that meaningful information may be conveyed by the order and duration of phrase categories, yet this has not been extensively examined in part because pertinent statistical tools are not available to extract this information. In the present study, we address this gap in our knowledge by developing and employing novel statistical tools and using them to explore the extent to which the duration and ordering of phrase categories

differ in the songs of male starlings singing to attract females in a spring (breeding) condition compared to the songs of males singing in an autumn (nonbreeding) condition.

We additionally explored the utility of these statistical methods by using them to shed light on how the brain fine-tunes and adjusts vocal behaviour seasonally. The medial preoptic nucleus (POM) is a brain area that is implicated in sexual motivation across vertebrate taxa, including birds (Balthazart, Absil, Gerard, Appeltants, & Ball, 1998; Balthazart & Surlemont, 1990; Lesions to the POM in male starlings during the breeding season strongly disrupt the production of sexually motivated courtship song (Alger, Maasch, & Riters, 2009; Alger & Riters, 2006; Riters & Ball, 1999) and implicate this area in song structure. Males with POM lesions sing shorter songs, and females approach POM-lesioned males less often than control males (Alger et al., 2009). These studies suggest that POM lesions make song less attractive to females, yet analyses have not been performed to examine the effects of POM lesions on the detailed structure of song. It has been proposed that lesions to the POM remove the sexual motivation to sing. Thus, POM-lesioned birds in spring condition are predicted to produce more autumn-like songs (i.e. songs that are not sexually motivated). Here we use the novel statistical techniques developed for comparison of spring and autumn song to test this prediction.

We recorded singing behaviour of male European starlings in two contexts: (1) males in photoperiodic and hormonal conditions mimicking the spring breeding season and in the presence of a female conspecific; and (2) males in conditions mimicking the autumn nonbreeding season in a social all-male flock. We also recorded singing behaviour of males in spring condition with bilateral POM lesions in the presence of a female conspecific. We examined differences in song structure among the three groups using semi-Markov methods, Bayesian estimation and classification trees.

METHODS

Animals

We captured 58 male and 48 female adult European starlings between December 2006 and February 2007 in Madison, Wisconsin, U.S.A. The birds were housed indoors in same-sex groups in stainless-steel cages on an 18:6 h light:dark cycle to induce a photorefractory state, a condition characterized by regressed gonads and a lack of sexual behaviour (Falk & Gwinner, 1988). All experiments were in accordance with the National Institutes of Health Guidelines and approved by the Institutional Animal Care and Use Committee of the University of Wisconsin (protocol code L00366).

Hormone Manipulations

Males were induced into one of two hormonal states, photostimulated (spring-like) and photorefractory (autumn-like), for song recording. We induced a spring-like state in 46 of the 58 males via direct hormonal and photoperiod manipulations. These males were gonadectomized and subsequently implanted with testosterone (T) so that POM lesion effects could be examined independent of any involvement of the POM in the regulation of T. Briefly, males were anaesthetized with isoflurane, a small incision was made just anterior

to the last rib, and gonads were removed using forceps (Alger et al., 2009; Alger & Riters, 2006). The incisions were sutured and males were allowed to recover on a heating pad. These males were then placed on an LD 6:8 h photoperiod for at least 8 weeks to induce photosensitivity, a photoperiod under which birds will become fully reproductively active in response to T but will not enter a state of photorefractoriness (Falk & Gwinner, 1988). These gonadectomized subjects received two subcutaneous 14 mm Silastic implants (1.47 mm inner diameter, 1.96 mm outer diameter; Dow Corning, Midland, MI, U.S.A.) filled with 10 mm of crystalline T (Sigma, St Louis, MO, U.S.A.) at the time of lesion surgery and were then placed on an LD 11:13 h photoperiod when they were moved into observation rooms for song recording (detailed below). Past work has shown that these photoperiods and T treatment maximize the likelihood that castrated male starlings will display song in a sexually relevant context in captivity (Alger & Riters, 2006; Riters & Ball, 1999) and result in circulating T concentrations that are within the physiological range of breeding males for at least 8 weeks (Duffy, Bentley, Drazen, & Ball, 2000). All birds recovered from gonadectomy and implant surgery without complications and were active and alert within minutes postsurgery.

The other 12 males were assigned to the autumn group and were housed on an LD 16:8 h photoperiod for at least 8 weeks, a photoperiod that induces and maintains photorefractoriness, an autumn-like condition characterized by regressed gonads and a lack of sexual behaviour. These males were not gonadectomized or implanted with hormones.

As in past studies from our laboratory (e.g. Alger et al., 2009; Alger & Riters, 2006), females to be used as behavioural stimuli were housed on an LD 11:13 h photoperiod for at least 12 weeks. These females received subcutaneous oestradiol-containing implants to facilitate sexual interest in males and male song production. Each female received two 18 mm Silastic implants (1.47 mm inner diameter, 1.96 mm outer diameter, Dow Corning) filled with 13 mm of 17-beta-oestradiol (Sigma). The hormone implantation procedure was identical to that used for male T implants.

Lesion Procedure

Spring-like males were randomly assigned to the spring POM-lesioned group ($N = 30$) or the spring control group ($N = 16$). We assigned more birds to be lesioned than to receive a sham operation because we expected that, due to natural variation in brain size among wild-caught starlings, some of the lesions would miss the POM. Higher numbers of POM-lesioned birds were also included because POM-lesioned animals are less likely to produce songs given the known role of this brain structure in courtship song (Alger et al., 2009; Alger & Riters, 2006; Riters & Ball, 1999). (Indeed in the present study only seven of the successful bilaterally POM-lesioned birds sang; see Results.)

Stereotaxic procedures were similar to those reported previously (Alger et al., 2009; Alger & Riters, 2006). Briefly, the anterior–posterior and ventral–dorsal zero coordinates were taken directly from the ear bar of a small animal stereotaxic apparatus (Kopf, Tujunga, CA, U.S.A.) fitted with a beak cone for gas anaesthesia. Males, anaesthetized with isoflurane, were placed into the stereotaxic apparatus with the beak angled at a 45° angle below the horizontal plane. The skull was opened above the POM and the lateral–medial zero

coordinate was taken from the midline at the surface of the brain. Lesions were made using ibotenic acid, an NMDA glutamate receptor agonist that induces apoptosis in target cells and is commonly used to lesion target tissue while sparing fibres of passage (Jarrard, 1989). Ibotenic acid (Sigma; 10 μg/μl 0.1 M sterile phosphate buffered saline (PBS), pH 7.4) was injected with a single bilateral injection of 0.2 μl into the POM with a 1 μl Hamilton syringe, extending from the zero coordinate to coordinates A (−0.5 mm), L (±0.15 mm) and V (7.0 mm). Spring control males received bilateral injections of sterile PBS, pH 7.4, into the same coordinates. The pipettes were left in situ for 5 min after each lesion, the skulls were closed with dental cement and the skin was sutured. Males were allowed to recover on a heating pad and, once recovered (typically within 1 h), they were returned to a clean cage and placed on an LD 11:13 h photoperiod. Birds recovered fully from lesion surgery without complications.

Song Recording and Behavioural Observations

Eleven days after lesion and implant surgeries, spring control and spring lesioned males were each introduced into a separate ($76 \times 76 \times 81$ cm^3) cage alone in a sound-attenuated indoor observation room on an LD 11:13 h photoperiod. Each cage contained a nestbox, two perches, and food and water ad libitum. A Shure SM57 microphone (Chicago, IL, U.S.A.) and an extra perch were located in the room, but outside of the cage. During each recording session, an observer released a stimulus female into the room, but outside of the cage, and recorded songs from behind a one-way mirror using a laptop computer, an M-Audio MoblePre pre-amplifier (Irwindale, CA, U.S.A.) and Avisoft recorder software (Berlin, Germany), an automated sound-recording software. Avisoft software was set to record and save all vocalizations longer than 10 s. Male song was recorded in multiple recording sessions over a 4-day period until 10 h of recording had been completed. After each recording session, the stimulus female was placed in the cage with the male for 10 min to encourage future singing behaviour. Most males did not sing when the female was placed in the cage, as in Riters and Ball (1999). Following song recording, we observed all spring control and spring lesioned birds for 20 min in the presence of a stimulus female. During this time, we recorded bouts of sexually relevant male behaviours (interactions with the nestbox and interactions with nest material) and other behaviours (feeding, drinking and preening).

We introduced groups of three autumn males, each wearing identifying coloured leg bands, into cage set-ups similar to those used to record songs of spring males. All three males were placed in the same cage and song was recorded from a single focal bird at a time (when no other birds were singing simultaneously). Autumn males did not receive a nestbox and female stimuli were not used to elicit song from males. For each autumn male, we recorded songs as previously described for spring birds above, often in multiple recording sessions, until at least 20 songs had been recorded or until 10 h of recording had been completed, whichever came first. Autumn birds were not behaviourally tested because males in autumn condition do not show sexual behaviour or sing in response to a female (e.g. Riters et al., 2000).

Histology

Subject males were euthanized via rapid decapitation. Brains were dissected from the skull and fixed overnight in a 5% acrolein solution. Males were checked at this time to confirm the presence of hormone implants and to check for testicular remnants. The brains were cryoprotected in 30% sucrose for 2 days, frozen with crushed dry ice and stored at −80 °C. Frozen brains were cut using a cryostat into 40 μm coronal sections and collected into PBS. All sections containing POM (i.e. sections beginning at the tractus septomesencephalicus to the anterior commissure) were Nissl stained with thionin.

We quantified the percentage of lesion damage located within and outside the POM for each bird from the Nissl-stained tissue using a Spot camera (Diagnostic Instruments, Sterling Heights, MI, U.S.A.) attached to a Nikon microscope (Melville, NY, U.S.A.) and a computer. As in our previous work (Alger et al., 2009; Alger & Riters, 2006), we acquired images at 20× magnification and traced the volumes of POM and the lesions inside and outside POM with a computer mouse using MetaVue software (Universal Imaging Corp., Downingtown, PA, U.S.A.). We reconstructed lesions by tracing the areas that showed obvious lesion damage, as confirmed using 100× magnification. We estimated POM and lesion volumes by multiplying the area for each section by 0.04 mm for the section thickness and summing these volumes for the length of the nucleus.

Of the 30 spring males to undergo POM lesion surgery, 22 had successful bilateral POM damage, six had unilateral POM damage and two had lesions completely outside of the POM. Both birds in which lesions missed the POM had very small lesions (<0.014 mm^3) in the nucleus commissurae pallii (nCPa). Sixteen males had successful sham surgeries. Consistent with a past study (Alger et al., 2009), the behaviour of the birds in which the lesions completely missed the POM were within the range of the sham control males, and the behaviour of the birds with unilateral POM lesions were highly variable. Thus, for the present study, males with complete misses of the POM were included in the control group and unilaterally lesioned males were excluded, resulting in 22 males in the spring lesioned group and 18 males in the spring control group.

Song Analysis

Only seven of the 22 spring lesioned males produced song, although one of them only produced a single song in 10 h of recording. Thirteen of the 18 spring control males produced song, although one of them only produced a single song in 10 h of recording. All 12 autumn males produced multiple recorded songs. We analysed song spectrograms (Fig. 1) with SASLab Pro Software (Avisoft). We operationally defined songs as vocalizations lasting at least 10 s with no pause longer than 2 s (as in Alger et al., 2009).

A single observer, blind to conditions, divided each song into four distinct phrase categories: whistles, variable phrases, clicks and trills (Adret-Hausberger & Jenkins, 1988; Chaiken, 1983; Chaiken, Gentner, & Hulse, 1997; Eens, 1997; Eens et al., 1991; Hausberger, 1997; Fig. 1). Whistles are long, low-pitched, relatively pure-tone phrases, at least some part of which falls below 5 kHz, with a minimum duration of 250 ms, separated with pauses of 1 s or more. Complex variable phrases are of low amplitude with short or no pauses between

them. Heterospecific imitations are often incorporated into complex variable phrases. Complex click phrases are characterized by the presence of a rapid series of clicks, called click trains, with maximum energy below 4 kHz and are delivered at a rate of about 10–16/s (Eens, 1997). Complex click phrases often include broadband buzzy and/or high-frequency elements superimposed on the click trains with no distinct temporal gaps between them. High-frequency trill phrases contain relatively pure-tone trills and remain above 6 kHz for at least 250 ms and often include rapid frequency modulations. These phrase types are the loudest in the typical song bout sequence and tend to be repeated more often than phrases of the previous categories. Typically, starling song phrases are produced in order from whistles, to variable phrases, to clicks, to trills; however, phrases can be produced in any order and a song may not include all phrase categories.

For each song, we recorded the time of each transition from one category to another (including the starting and ending time for the song).

Rationale for statistical approach

Common measurements taken from sonograms relate to durations, frequencies (i.e. pitch), repetition rates and repertoire sizes of song elements such as notes, syllables, phrases, phrase types and songs. Many of these attributes can be compared statistically across groups with common statistical techniques such as ANOVAs or t tests, or their nonparametric counterparts (e.g. Catchpole, 1976). Additionally, measures of within-individual variation, such as the standard deviations of previously mentioned measures, can be compared using these statistical techniques to reflect differences in stereotypy. However, information contained in the order of song elements within an individual's song and the tendency for an individual to reproduce this order across songs (an important measure of stereotypy) is more complicated and cannot be as easily extracted statistically.

One commonly used statistical tool for the analysis of ordered data is Markov models. A Markov chain is a stochastic process that has the Markov property, meaning that future states depend only on the present state and are thus independent of past states. In this case, we can think of a starling song as a chain of states, where the states are the song elements being sung at each moment in time, plus a 'stop' state indicating the end of the song. At each step in the chain, the system may change from its current state to another state (referred to as 'transitions'), and there is a particular probability associated with each possible transition (referred to as 'transition probabilities'). Markov chains can have variable but finite memory length; a Markov chain with a memory of length m is said to be an mth-order Markov model. Thus, in a first-order Markov model, the next state can be predicted given only the current state. In a second-order Markov model, the combination of the current state and the state immediately preceding it are needed to reliably predict the next state. There are statistical tests that can assess the order of the Markov chain that can best predict future states within the chain, and this order can be used as a measure of song stereotypy. This technique has been used to analyse vocalizations from multiple species, including songbirds (e.g. Martin, 1990), hummingbirds (e.g. Rusch, Pytte, & Ficken, 1996; Rusch, Thusius, & Ficken, 2001) and whales (e.g. Miksis-Olds, Buck, Noad, Cato, & Stokes, 2008).

Markov models, as they have traditionally been applied to the study of birdsong and other sequenced behaviours, have many drawbacks. One drawback of this technique is that the order of a traditional Markov model applies to all states, even if not all states predict the next state with the same degree of power. Additionally, biologically meaningful interpretations become more difficult with increasing order. A third drawback is that the type of Markov models used thus far in the analysis of vocalizations have been discrete models, which do not make use of duration data available in spectrogram recordings. Markov models that do make use of duration information require a specific distribution of those durations that make them intractable for the study of birdsong. Furthermore, the maximum likelihood methods that are commonly used to determine Markovian transition probabilities are heavily biased by rare occurrences and small sample sizes (see Supplementary Material 1). These shortcomings are not inconsequential. In fact, recent analyses suggest that traditional Markov methods do not reflect the vocal systems of many animal species as well as was previously assumed (Jin & Kozhevnikov, 2011; Kershenbaum et al., 2014). To address these issues, we apply novel semi-Markov analyses (which do not have distribution requirements for element durations), Bayesian inference (which can be more reliable in the prediction of rare events) and classification trees to the analysis of starling song (for more detail on the conceptual rationale for these techniques, see Supplementary Material 1).

Statistical model of birdsongs

Each birdsong was modelled as the realization of a semi-Markov process, beginning with an artificial start state, proceeding through a sequence of phrases of random duration and ending with an artificial stop state. A semi-Markov process is a type of Markov model in which the process that determines the transition order is a Markov chain and the durations of each state are random variables with any distribution. (In a Markov model, durations must be exponentially distributed, which our duration data were not.)

The first step in a semi-Markov analysis is defining the state space. One vital assumption of semi-Markov processes is that the distribution of the durations within each state is independent of the realized states and the durations at other parts of the sequence. However, graphical assessment and Wilcoxon two-sample testing of durations of whistle phrases showed marked differences in median duration depending on where the whistle phrase appeared in the song. The other three phrase categories showed no such significant duration differences. Thus, we divided whistle phrases into three distinct states as follows: whistle1, a whistle phrase that does not end a song; whistle2, a whistle phrase that ends a song but does not start it; and whistleSong, a single whistle phrase that comprises an entire song (it both starts and ends a song). The resulting state space for the semi-Markov process contained eight mutually exclusive states (Start, whistle1, whistle2, whistleSong, variable phrases, clicks, trills, Stop). Start and Stop had no duration, but each of the other states had a corresponding phrase duration distribution.

The next step in a semi-Markov analysis is characterizing the transition probabilities. We used Bayesian methods (see Supplementary Material 2) to estimate a matrix of transition probabilities separately for each bird. Each bird's set of transition probabilities can be characterized by the seven probability distributions of leaving each state (Start, whistle1,

whistle2, whistleSong, variable phrases, clicks, trills). Each of these states can transition to whistle1, whistle2, whistleSong, variable phrases, clicks, trills and Stop with an unknown probability greater than 0 and less than 1, with the following limitations: (1) a state cannot transition to itself and whistle states cannot transition to other whistle states; (2) Start cannot transition to whistle2 or Stop; (3) Start is the only state that can transition to whistleSong; (4) whistle2 and whistleSong must transition to Stop; and (5) no state can transition to Start. Thus, the state definitions constrain several transitions to be either 0 (i.e. impossible) or 1 (i.e. required). The *k* unconstrained transition probabilities are the parameters to estimate, subject to each row summing to one (see Table 1).

The commonly used frequentist (or maximum likelihood) paradigm presents the problem here that any transitions that a given bird does not display would be assigned a probability of 0, and likewise, any transitions that a given bird always displays would be assigned a probability of 1, both of which overestimate our certainty. In contrast, within the Bayesian paradigm, a prior distribution (i.e. our belief about the unknown values before experimental data is seen) is updated by the data to create a posterior distribution. If we estimate transition probabilities from the resulting posterior distributions, the resulting estimates may approach, but will never equal, 0 or 1. Thus, we used Bayesian estimation to estimate the *k* transitions from each state that were not defined to be 0 or 1.

Each set of transitions from a given state follows an unknown multinomial distribution. For a set of transitions from a given state, let *k* be the number of states (indexed by *j*), *y* be the vector of counts of the number of transitions to each state, and θ be the vector of probabilities of each associated transition. The likelihood (i.e. Pr(data|θ)) can be described by: Pr($y|\theta$) is proportional to $\prod_{(j=1,...,k)} \theta_j^{y_j}$ where $\sum_{(j=1,...,k)} \theta_j = 1$. There are various options for the choice of prior distribution (i.e. Pr(θ)). The natural conjugate prior distribution (i.e. the set of densities with the same form as the likelihood that result in a posterior distribution of the same form) is the Dirichlet distribution with parameters $a = a_1, a_2, ..., a_k$. The Dirichlet distribution has a density where Pr($\theta|a$) is proportional to $\prod_{(j=1,...,k)} \theta_j^{y_j-1}$ and $_{(j=1,...,k)} \theta_j = 1$. To diminish the effect of our selected prior distribution, we set $_{(j=1,...,k)} (a_j) = 1$, which means that the prior distribution had about as much influence on the posterior distribution as a single song. We initially used a uniform prior distribution, such that $a_j = 1/$(number of possible transitions from state *j*). We then repeated the analysis with a prior distribution weighted by our previous knowledge that the most common starling song structure is whistle–variable phrase–clicks–trills. In this case, the most common transitions (i.e. Start→whistle, whistle→variable phrase, variable phrase→clicks, clicks→trills and trills→Stop) were assigned $a_j = \frac{1}{2}$, whereas all other possible transitions were assigned $a_j = 1/(2 \times$ (number of possible transitions from state *j* − 1)). Because the results using both prior distribution selections were similar and the weighted prior may reinforce reliability in cases of small sample sizes, only the results of the weighted prior distribution are reported below.

Calculation of distances between pairs of birds using transition data

We defined the transition distance between each pair of birds as the sum of the total variation distances between their corresponding nontrivial transition probability distributions

(i.e. all transition probabilities not equal to 0 or 1). Total variation distance between two probability measures can be thought of as the largest possible difference between the two probabilities assigned to some set of possible values. If $f_1(x)$ and $f_2(x)$ are the posterior probability distributions of transitioning from one particular state to all other states for two different birds, then the total variation distance between these distributions can be calculated by $d(f_1(x), f_2(x)) = (1/2) \sum_{(x)} | f_1(x) - f_2(x)|$.

To account for uncertainty in these transition distributions as described by their posterior distributions, for each pair of birds, for each of the five nontrivial probability distributions, a realization of the transition probabilities was drawn from each bird's respective posterior Dirichlet distribution and their total variation distance was calculated. This process was repeated for 10 000 trials and the resulting total variation differences were averaged over the trials, resulting in five total variation distances, each ranging from 0 to 1, for each pair of birds. The sum of these numbers is the calculated transition distance between the pair of birds and it ranges from 0 to 5.

Bird characterization and calculation of distances between pairs of birds using duration data

We modelled the log-transformed durations of each phrase category as normally distributed. Normality was assessed with visual inspections of density plots and Kolmogorov–Smirnov tests. Each normal distribution is characterized by its mean and variance, which we estimated with Bayesian inference. For the duration of a given state, let $y = \{ Y_1, Y_2, \ldots, Y_n\}$ be the n observed durations of that state by a given bird and each set of observations $y|\mu,\sigma^2$ is independent and identically distributed normally with mean μ and variance σ^2. Because parameters μ and σ^2 are unknown, they also have distributions. The prior distributions of μ and σ^2 are: $\sigma^2 \sim$ inverse gamma($v_0/2$, $v_0\sigma^2_0/2$) and $\mu|\sigma^2 \sim$ normal(μ_0, σ^2/κ_0), where $v_0\,\sigma^2_0$, μ_0 and κ_0 are fixed parameters that determine the means, variances and shapes of these distributions. The posterior distributions of μ and σ^2 are: $\sigma^2 \sim$ inverse gamma($v_n/2$, $v_n\sigma^2_n/2$) and $\mu|\sigma^2 \sim$ normal(μ_n, σ^2_n/κ_n), where v_n, σ^2_n, μ_n, and κ_n are parameters determined by the values of the prior parameters and the data according to the following formulas:

$$
\begin{aligned}
v_n &= v_0 + n \\
\kappa_n &= \kappa_0 + n \\
\mu_n &= (\mu_0\kappa_0 + n\,(y-\text{bar})) \,/\, \kappa_n \\
\sigma^2_n &= \left(v_0\sigma^2_0 + (n-1)\sigma-\text{hat}^2 + \left((y-\text{bar} - \mu_0)^2 \times (n\,\kappa_0/\,\kappa_n)\right)\right) \,/\, v_n
\end{aligned}
$$

where y-bar is the sample mean and σ-hat^2 is the sample variance. Thus, for each bird, the estimated variance was randomly drawn from an inverse gamma distribution with shape parameter $v_n/2$ and scale parameter $v_n\sigma^2_n/2$, and the estimated mean was randomly drawn from a normal distribution with mean μ_n and variance σ^2_n/κ_n. Prior distributions for the log durations of each state were determined by informed elicitation, a process of extracting expert knowledge about an unknown quantity when published data are not available. In this process, we estimated the first and third quantiles for each state's duration ('lo' and 'hi', respectively) and used this information to characterize the associated normal distribution. The prior mean (μ_0) can be found by $\mu_0 = (\log(\text{hi}) + \log(\text{lo})) / 2$, and the prior standard

deviation (σ_0) can be found by $\sigma_0 = (\log(hi) - \log(lo)) / (2 \times \Phi\ (0.75))$. We then generated random data from a normal distribution with mean μ_0 and variance σ^2_0 and confirmed with box plots whether the quantiles seemed reasonable, with the intention of producing distributions with wider variance than the true sample distributions. The values for the prior distributions for all states except for variable phrases were: $\upsilon_0 = 4$, $\kappa_0 = 1$, $\mu_0 = 1.589$, $\sigma^2_0 = 1.764$ (which correspond to 80% of the prior durations of these categories falling between 0.9 and 26.8 s). The values used for the prior distribution of variable phrases were: $\upsilon_0 = 4$, $\kappa_0 = 1$, $\mu_0 = 3.198$, $\sigma^2_0 = 1.423$ (which correspond to 80% of the prior durations of this category falling between 5.3 and 113.0 s). These distribution parameters were chosen to have larger than expected variance (for more information on this statistical approach, see van der Vaart, 2000).

We defined the duration distance between each pair of birds as the sum of total variation distances between the corresponding distributions of log durations. As for the transition distance, we accounted for posterior uncertainty in these distributions by sampling from the posterior distributions many times and averaging the corresponding distances. Thus, for each pair of birds, for each of the six phrase category states, a posterior mean was drawn from each bird's respective normal distribution and a posterior variance was drawn from each bird's respective inverse gamma distribution and the total variation distance between the two resulting normal curves was calculated. Specifically, if $g_1(x) \sim N(\mu_1, \sigma^2_1)$ and $g_2(x) \sim N(\mu_2, \sigma^2_2)$ are the probability density functions of log durations for one particular state for two different birds, then the total variation distance can be thought of as the total area under one normal curve that is not under the other normal curve. This can be calculated by finding the points where the two curves intersect (for simplicity, call these points A and B) and calculating the difference between the areas under these curves between these points:

$$
\begin{aligned}
&d\left(g_1\left(x\right),\quad g_2\left(x\right)\right)\\
&\quad = \mid\ \left(\Phi\left(A,\quad \mu_1, \sigma^2_1\right)\quad -\quad \Phi\quad \left(B, \mu_1,\quad \sigma^2_1\right)\right)\\
&\quad -\ \left(\Phi\quad \left(A,\quad \mu_2,\quad \sigma^2_2\right)\quad -\quad \Phi\quad \left(B,\quad \mu_2,\quad \sigma^2_2\right)\right)\mid\ \text{for}\quad \sigma^2_1 <\quad \sigma^2_2.
\end{aligned}
$$

Again, the areas below $g_1(x)$ and $g_2(x)$ are both 1, so the distance score between them ranges from 0 to 1 (Fig. 2). For each pair of birds, this process was repeated for 10 000 trials for each of the six phrase category states and the resulting total variation differences were averaged over the trials. The sum of these numbers, which can range from 0 to 6, is the calculated duration distance.

Analysis of distances between groups of birds

We generated a distance matrix by summing the distance with respect to the transition probabilities and the distance with respect to the durations for each pair of birds. If spring song is consistently different from autumn song, then we should see smaller distance scores when comparing within groups as the birds were currently assigned than we would see if the birds were assigned 'spring' and 'autumn' labels at random. We conducted a permutation test to test this hypothesis. Using the distance scores from the 12 spring control birds that sang multiple songs and the 12 autumn birds, we calculated and summed the average within-

group distance scores. We then permuted the 12 spring labels and the 12 autumn labels at random, each time generating a distance score table and the corresponding sum of mean within-group distance scores for 10 000 runs. The *P* value was generated by comparing the true value of the sum of mean within-group distance scores to the normally distributed, randomly permuted data. This process was repeated for a comparison of song produced by spring control birds and spring lesioned birds, and also for a comparison of song produced by autumn birds and spring lesioned birds. Bonferroni corrections were used when interpreting *P* values.

We also conducted distance matrices and corresponding permutation tests for all six phrase categories separately to assess differences between groups that may be specific to individual phrase categories. For each phrase category, the difference score assigned to a pair of birds reflected differences in transition probability distributions and in the distribution of log durations for that phrase category alone, resulting in a distance score that ranged from 0 to 2.

Discrimination of spring and autumn songs with classification trees

We used classification trees to assess the ability of each song structure variable to differentiate songs of spring control males and autumn males to develop descriptive rules to which we could compare POM-lesioned male songs. In the simplest form of classification tree, each binary split of the predictor space Z relates to a single variable and the split is in the form of a rule that divides the cases into two groups, $\{Z < \gamma\}$ and $\{Z \geq \gamma\}$. The probability distribution of the k classes for node i of the tree is p_{ik}, and each possible split yields a random sample n_{ik} of cases from the multinomial p_{ik} on each side of the node. The conditional likelihood is proportional to $\prod_{(\text{leaves } i)} \prod_{(\text{classes } k)} p_{ik}{}^{n_{ik}}$. The deviance for node i is $D_i = -2 \sum_k n_{ik} \log p_{ik}$, and the total deviance of the tree is the sum over all of the leaves $D = \sum_i D_i$. The deviance is evaluated for all possible binary splits of the predictor variable. The 'best split' at a node is defined as the split that minimizes the deviance (for more information on these statistics, see Hill & Pawel, 2006; Kvam & Vidakovic, 2007).

We entered the numerical data for each song structure variable and the categorical assignment of spring control or autumn for all spring control and autumn birds that produced at least two songs (spring control: $N = 12$; autumn: $N = 12$) into the function tree in the R package 'tree' (version 1.0-28) and extracted the deviance and best split for the initial node. The song structure variables included the Bayesian posterior transition probabilities, mean log durations and standard deviations of log durations for each phrase category. We used a permutation test to determine statistically significant initial node deviance scores. Briefly, we permuted treatment labels (spring control and autumn) over all birds and calculated the initial node deviances for all song structure variables for 5000 trials. We considered initial node deviances that were less than 5% of the deviances calculated by permutation to be statistically significant; initial node deviances that were less than 10% of the deviances calculated by permutation were considered a statistical trend. We then used extracted rules for the best split of the initial nodes to assess song produced by POM-lesioned males to determine whether POM lesions caused songs to be more autumn-like.

A novel measure of stereotypy

Each song can be characterized as one would characterize a bird that produced only that one song. Similarly, each song produced by a single bird can be compared to all other songs produced by that bird and a distance matrix of pairwise distances can be generated for each bird. When all pairwise distances for a single bird are averaged, the result is a number (called d) that ranges from 0 to $(2s + 1)$, where s is the number of states, and is high for birds with high song-to-song variability and low for birds with low song-to-song variability. Stereotypy is a high degree of consistency between songs produced by a single bird (i.e. low song-to-song variability). Thus, we calculated a 'stereotypy score' for each bird that sang more than one song as $1/(d+1)$, which ranges from near zero (for a bird with maximally different songs) to 1 (for a bird with identical songs).

Statistical software and tests

We performed all statistical analyses using R v.2.10.1 statistical software (R Core Team, 2009). We compared differences between groups in behaviour counts, phrase category durations and phrase category occurrences using ANOVAs with Tukey HSD post hoc tests, using log transformations where appropriate.

RESULTS

Initial Exploratory Statistical Analysis and Definition of State Space

The song recording collection consisted of songs recorded from spring control males ($N = 13$ birds, 1–82 songs per individual), spring POM-lesioned males ($N = 7$ birds, 1–42 songs per individual) and autumn males ($N = 12$ birds, 10–63 songs per individual).

Whistle1 had a significantly longer duration than whistle2 (mean ± SD: whistle1: 7.30 ± 2.38 s; whistle2: 3.86 ± 2.43 s; Wilcoxon two-sample test: $W = 46{,}037.5$, $P < 0.001$). The duration of whistleSong (17.68 ± 6.64 s) was significantly longer than that of whistle1 ($W = 3{,}642$, $P < 2.2\text{e-}16$) and whistle2 ($W = 122$, $P < 0.001$). The increased duration of whistleSong was at least in part due to the definition of a song being a vocalization of at least 10 s, which set a lower-bound for any phrase categories that both started and ended a song. Durations of variable phrases (13.96 ± 4.60 s), clicks (8.05 ± 2.49 s) and trills (6.20 ± 2.30 s) were consistent without respect to their location within the song.

Parametric Group Comparisons for Song Measures

Lesioned birds tended to produce fewer songs (mean ± SD: 11.83 ± 14.86 songs) than control birds (37.25 ± 27.23 songs) per 10 h recording session (Student's t test: $t_{16} = 2.11$, $P = 0.051$; Fig. 3a). Song bout length was log-normally distributed, so ANOVA tests followed by Tukey HSD post hoc tests were appropriate for log-transformed duration data (means ± SD in the text and figures indicate original, untransformed, duration values). The log(song length) was significantly shorter for autumn birds (28.97 ± 3.81 s) than for spring control birds (39.47 ± 7.59s) and spring lesioned birds (35.54 ± 3.79 s) (ANOVA: $F_{2,27} = 10.18$, $P < 0.001$; Tukey HSD post hoc tests: autumn versus spring control: $P < 0.001$; autumn versus spring lesioned: $P = 0.045$; Fig. 3b).

For each bird, we calculated the proportion of songs that contained each phrase category, the average number of repetitions of each phrase category per song and the average duration of each phrase category per song. All phrase categories were consistent with log-normal duration distributions, so ANOVA tests followed by relevant Tukey HSD post hoc tests were appropriate for log-transformed data. Because log(song length) differed between groups, we also compared the log(phrase category duration) multiplied by a correction factor of the mean song bout length across all birds divided by the mean song bout length for the group. This allowed us to compare phrase category duration among groups independent of differences in overall song length. Songs produced by autumn birds contained a significantly higher proportion of whistle2 phrases (mean $\pm$ SD = 0.10 ± 0.06) than songs produced by spring control birds (0.05 ± 0.05) and spring lesioned birds (0.004 ± 0.01) ($F_{2,27} = 8.51$, $P = 0.001$; autumn versus spring control: $P = 0.026$; autumn versus spring lesioned: $P = 0.002$; Fig. 4a). Autumn birds also produced significantly longer whistle2 phrases per song (mean $\pm$ SD = 4.41 ± 2.81 s) than spring control birds (3.36 ± 1.61 s) and the one spring lesioned bird in which it occurred (1.21 s), and the spring control birds produced longer whistle2 phrases per song than the spring lesioned bird ($F_{2,27} = 12.85$, $P < 0.001$; autumn versus spring control: $P = 0.024$; Fig. 4b). This result was also true for corrected log-transformed whistle2 durations (autumn: 5.21 ± 3.31 s; spring control: 2.91 ± 1.39 s; spring lesioned: 1.16; $F_{2,27} = 12.85$, $P < 0.001$; autumn versus control: $P = 0.008$). The log durations of variable phrases were significantly longer in spring control birds (15.69 ± 3.63 s) than in autumn birds (11.86 ± 4.36 s) ($F_{2,27} = 3.55$, $P = 0.043$; autumn versus spring control: $P = 0.036$; Fig. 4c), but this difference was eliminated by correcting for group differences in overall song bout length. The log durations of variable phrases also tended to be longer in spring control birds than in spring lesioned birds (14.69 ± 5.76 s). The log durations of trills were significantly longer in spring control birds (7.64 ± 2.07 s) than in autumn birds (4.73 ± 1.75 s) and tended to be longer in spring lesioned birds (6.28 ± 2.08 s) than in autumn birds ($F_{2,27} = 7.12$, $P = 0.003$; autumn versus spring control: $P = 0.003$; autumn versus spring lesioned: $P = 0.065$; Fig. 4d). The significant difference between spring control birds and autumn birds also held for corrected log durations of trills (spring control: 6.62 ± 1.79 s; spring lesioned: 6.05 ± 2.00 s; autumn: 5.59 ± 2.07 s; $F_{2,27} = 3.72$, $P = 0.037$; autumn versus control: $P = 0.037$). No other differences in the proportion of songs with each phrase category, number of repetition of phrase categories or average phrase category log duration were found.

Song Structure Distance Analysis

The permutation test confirmed that song data from spring control birds and autumn birds were separate groups ($P = 0.003$). Permutation tests also showed that song produced by spring POM-lesioned males differed significantly from song produced by autumn males ($P = 0.005$) but not from song produced by spring control males ($P = 0.115$).

We also used permutation tests to compare transition and duration information separately for each of the six phrase categories. These tests revealed that spring control birds showed significant differences in variable phrases ($P < 0.001$) and trills ($P < 0.001$) compared to autumn males. Lesioned birds showed significantly different variable phrases ($P = 0.010$) than spring control males and significantly different whistle2 ($P < 0.001$) and variable phrases ($P = 0.007$) than autumn males.

Classification Tree Analysis

Spring control male song and autumn male song could be significantly distinguished by the mean duration of trills (dev = 17.51, split = 1.61, $P < 0.05$), mean duration of variable phrases (dev = 19.87, split = 2.31, $P < 0.05$), mean duration of clicks (dev = 22.91, split = 2.13, $P < 0.05$), mean standard deviation of clicks (dev = 22.91, split = 0.56, $P < 0.05$), probability of transitioning from trills to whistle1 (dev = 25.01, split = 0.003, $P < 0.05$) and probability of transitioning from trills to whistle2 (dev = 25.01, split = 0.003, $P < 0.05$; Table 2). The probability of transitioning from clicks to whistle2 showed a trend of distinguishing spring control male song from autumn male song (dev = 25.12, split = 0.05, $P < 0.10$; Table 2).

Of the seven classification tree rules for distinguishing spring control from autumn male song, none were followed by all 12 spring control males, and four were not followed by any of the 12 autumn males (Table 2). Likewise, no bird's collection of songs followed all seven rules indicative of a spring bird. However, only spring control males followed at least five of the seven rules. Three autumn birds and no spring control birds followed none of the rules that were indicative of a spring bird.

The six spring birds with bilateral POM lesions that sang more than one song followed some of the rules indicative of spring song, but not others (Table 2). For example, POM-lesioned males were more likely to have a low probability of transitioning from clicks to whistle2 states, to have long trills and to have long variable phrases, all attributes characteristic of spring song. However, they were also more likely to have short and variable click phrases and high probabilities of transitioning from trills to whistle1 and whistle2, all attributes characteristic of autumn song.

Measure of Stereotypy and Other Behavioural Measures

Autumn birds had significantly less stereotyped songs (0.33 ± 0.02) than spring control birds (0.38 ± 0.05) and spring lesioned birds (0.41 ± 0.05) ($F_{2,26} = 7.08$, $P = 0.004$; autumn versus spring control: $P = 0.049$; autumn versus spring lesioned: $P = 0.004$; Fig. 5a). With respect to the transition orders alone, autumn birds produced less stereotyped songs (0.49 ± 0.04) than spring lesioned birds (0.57 ± 0.08) and tended to produce less stereotyped songs than spring control birds (0.53 ± 0.07) ($F_{2,26} = 3.93$, $P = 0.032$; autumn versus spring lesioned: $P = 0.031$; Fig. 5b). With respect to phrase category durations alone, autumn birds produced less stereotyped songs (0.51 ± 0.02) than both spring control birds (0.56 ± 0.04) and spring lesioned birds (0.59 ± 0.03) ($F_{2,26} = 12.81$, $P < 0.001$; autumn versus spring control: $P = 0.005$; autumn versus spring lesioned: $P < 0.001$; Fig. 5c).

We compared behavioural measures between spring control and spring lesioned groups. Among all control and lesioned birds tested, control males travelled more with nest material and interacted more with their nestboxes than did lesioned males (nest material: spring control: 0.35 ± 0.61; spring lesioned: 0.05 ± 0.22; $t_{36} = 2.15$, $P = 0.039$; nestbox interaction: spring control: 4.71 ± 5.07; spring lesioned: 1.90 ± 3.06; $t_{36} = 2.10$, $P = 0.042$). Also, lesioned males fed and drank more often than control males (feeding: spring control: 2.88 ± 2.91; spring lesioned: 5.10 ± 3.45; $t_{36} = 2.10$, $P = 0.042$; drinking: spring control: 2.53

± 1.84; spring lesioned: 4.71 ± 3.00; t_{36} = 2.62, P = 0.013). When we compared the behaviours of only control and lesioned birds that sang, control males interacted more with their nestboxes than did lesioned males (spring control: 6.42 ± 5.12; spring lesioned: 1.33 ± 2.16; t_{16} = 2.30, P = 0.035) but did not travel more with nest material. Singing lesioned males drank more often, but did not feed more often, than singing control males (spring control: 1.92 ± 1.78; spring lesioned: 5.00 ± 4.05; t_{16} = 2.28, P = 0.037).

DISCUSSION

We developed novel statistical tools to analyse sequence data that differ in both order and duration and applied them to sequences of birdsong. We first discuss what these statistical methods reveal about birdsong and then highlight the strengths of this statistical approach.

Differences in Stereotypy of Songs of Male Starlings in Spring and Autumn Condition

Some seasonally breeding male songbirds adjust song so that it is more stereotyped when used to attract females (Nottebohm et al., 1986; Smith, Brenowitz, & Wingfield, 1997). This had been suggested anecdotally for starlings (Adret-Hausberger & Jenkins, 1988); however, the only previous study that quantified stereotypy in starlings (using qualitative visual inspection of sonograms) revealed no seasonal differences (Van Hout et al., 2009). In contrast, using our stereotypy score, we found that song was more stereotyped in male starlings in spring condition than in autumn condition.

The stereotypy score that we used is, to our knowledge, the only score that incorporates information about both transition probabilities and duration distributions, and we show here that the two types of information can be individually assessed. There are no consistently used measures of stereotypy in studies of birdsong. Stereotypy ratings have been made using methods as simple as subjectively comparing pairs of syllables and denoting them 'more similar' or 'less similar' (e.g. zebra finches, *Taeniopygia guttata*: Jones, ten Cate, & Slater, 1996). However, this method is not as effective with species with more complex and variable songs such as starlings. Pairwise frequency-time cross-correlation values of digitized song elements have been used in studies in which note structure is the feature of interest (e.g. Smith, Brenowitz, Beecher, & Wingfield, 1997), but longer song elements such as phrases or phrase categories often introduce too much variability for this method to be useful. A stereotypy score for more complex song structure has been calculated using the average of two related measures of stereotypy: sequence linearity and sequence consistency (Scharff & Nottebohm, 1991), but this measure does not include duration information. Another method for assessing the stereotypy of note order uses the concept of transition entropy from Markov methods to describe stereotypy (Sakata, Hampton, & Brainard, 2008). This method analyses each transition type separately, and because transition entropy measures assume all elements in the sequence are independent, they are probably not suitable for quantifying the complexity of most birdsong (Kershenbaum, 2014).

The measure of stereotypy that we present allows individual assessment of transition probabilities and duration distributions. Furthermore, the use of Bayesian posterior means as estimates for transition probabilities provides the added benefit that no transition is assigned a probability of exactly 0 or 1 unless by the design of the definitions. A second advantage of

the Bayesian approach is that our simulation-based distance comparisons between birds naturally account for uncertainty due to variation in the number of songs each bird sang. Finally, an advantage of the present approach is that these concepts can be applied to any level of song element, including notes, syllables, phrases and phrase categories.

Additional Structural Differences in Songs of Starlings in Spring and Autumn Condition

Consistent with past studies (Riters et al., 2000) starling song produced in a spring context was longer than song produced in an autumn context. With respect to specific phrase categories, using traditional parametric statistical approaches (e.g. ANOVAs) and classification tree analysis we found that variable phrases, clicks and trills had longer durations and clicks had less variable durations in spring song than in autumn song. Furthermore, trills were longer in spring control males than in autumn males after correcting for overall differences in song length. These findings are consistent with a study in starlings that showed that males' repertoires of variable and click phrase types and phrase repetition rates for trills are larger in gonadally intact (sexually motivated) male starlings than in castrated (nonsexually motivated) male starlings in spring (Van Hout et al., 2009). Past studies have also shown that male starlings tend to produce more of these phrase types during courtship (Adret-Hausberger & Jenkins, 1988; Eens, Pinxten, & Verheyen, 1988; Eens, Pinxten, & Verheyen, 1993; Henry, Hausberger, & Jenkins, 1994; Verheyen, 1980).

The finding that male starlings invest in the production of variable, click and trill phrases in spring condition suggests that these phrases may convey meaning about male quality. Females tend to prefer males that are most capable of performing energetically costly acts (reviewed in Byers, Hebets, & Podos, 2010). With increased duration comes increased oxygen consumption (Franz & Goller, 2003), and increased complexity provides challenges to neural and motor systems. It has been proposed that variable phrases in starlings function in female attraction because they are typically the longest and most complex of all the phrase categories (Adret-Hausberger & Jenkins, 1988). Clicks may represent a motor trade-off between the rate of the click trains and the quality of superimposed notes. The production of high-frequency trills requires a significantly higher oxygen consumption rate than the production of other phrase types (Oberweger & Goller, 2001). The brains of female starlings are more responsive to these high-frequency trill phrases during the breeding season than outside of the breeding season (Cousillas et al., 2013). The present findings are thus consistent with the possibility that male starlings infuse song with more energetically costly elements at the time of year when females are selecting mates based on male motor performance, including vocal motor skills displayed in song (reviewed in Byers et al., 2010).

In contrast to variable, click and trill phrases, we found that whistle 2 occurred more often and at longer durations in song produced by autumn condition males that were singing in all-male flocks, compared to spring condition males that were singing to attract females. This finding is consistent with the proposed role of whistles in male–male interactions (Hausberger, 1997). However, previous results also suggest that some whistle types may be more common when a pair is formed in spring (Henry et al., 1994), and other results indicate that whistles either do not differ across contexts or may be used slightly more often during interactions with females during the breeding season (Eens et al., 1993). It has also been

proposed that distinct categories of whistles convey meaning about species or individual identity (Hausberger, 1997; Henry et al., 1994); however, we did not attempt to distinguish among distinct whistle types in the present study. Consistent with a role for at least some whistle types in mate attraction, an auditory-processing region in the female brain is more responsive (measured using fMRI activity) to species-specific male whistle types during the breeding season (De Groof, Poirier, George, Hausberger, & Van der Linden, 2013). Future studies in which distinct whistle types are categorized and the influence of breeding condition versus pairing status are needed to further understand this phrase category.

Effects of POM Lesions on Song Structure

In spring males, POM lesions significantly reduced the proportion of males to produce song and the number of songs produced by those who did sing. Although this was consistent with previous studies (Alger & Riters, 2006; Riters & Ball, 1999), the small sample sizes of both the number of POM-lesioned starlings who sang and the number of songs produced reduced our ability to identify conclusive effects of POM lesions on starling song structure. Despite these challenges, we did find that the duration of variable phrases in POM-lesioned birds was intermediate to that of spring control male song and autumn male song. Additionally, POM lesions had a profound effect on the whistle 2 category, in that only one lesioned bird produced this category, he did so only once, and it was significantly shorter in duration than those produced by either spring control males or autumn males. The classification tree analysis revealed that POM-lesioned males produced song that was spring-like in that it had long trills, long variable phrases and few whistle 2 phrases, but it was autumn-like in that it had short and variable click phrases and high probabilities of transitioning from trills to whistle phrases. The findings are consistent with a role for POM in adjusting structural aspects of song.

Assessment of Effectiveness of Statistical Methodology

Permutation tests revealed that the song structure of spring control and autumn birds differed in both transition probabilities (i.e. phrase category order) and phrase category durations. However, there was still overlap between the two sets of points, which is biologically intuitive. Data from some birds resulted in greater apparent distances from other birds due to small sample sizes. The Bayesian prior distributions were designed to have wide variances that would be reduced around the appropriate mean when updated with a bird's data. If a bird produced only a few songs, the posterior distribution would have a wide variance, which could result in greater distances from other birds than if the bird had sung more songs.

This method was also used in the analysis of data pertaining to individual phrase categories. Permutation tests indicated that spring control birds had significant differences in variable phrases and trills compared to autumn males, confirming the results seen using ANOVA tests. These tests also revealed that POM-lesioned birds significantly differed in the production of variable phrases compared to spring control males and in whistle 2 and variable phrases compared to autumn males, findings we were not able to detect using the ANOVA methods alone. These tests may have been more sensitive to these differences in part because the data used included both transition and duration data, as opposed to duration data alone.

Furthermore, the classification tree approach allowed for the creation of 'rules' that could be used to categorize a song or group of songs as spring-like or autumn-like with respect to potentially meaningful song attributes. This technique confirmed that the durations of variable phrases, trills and clicks, as well as several probabilities of transitioning to whistles, are important to differentiating spring-like and autumn-like song in starlings. However, this approach highlighted that POM-lesioned birds did not produce songs that were universally spring-like or autumn-like, providing novel insight to how the brain may regulate different song attributes.

Comparison to Other Sequence Analysis Methods

Markov chains and hidden Markov models (a generalization of the Markov model) are the most common models used to examine animal vocal sequences and they can do a reasonable job of capturing information about order, repetition and diversity of elements. However, they generally require long sequences of elements, large sample sizes and a stationary transition matrix, and they are limited in their utility for analysing temporal structure information (such as duration) (Kershenbaum et al., 2014; ten Cate, Lachlan, & Zuidema, 2013; ten Cate & Okanoya, 2012). Recent analyses have in fact demonstrated that many animal vocal sequences are not well described by Markov models (Jin & Kozhevnikov, 2011; Kershenbaum et al., 2014). Our combination of a semi-Markov model with Bayesian inference has the potential to be more reliable in the prediction of rare events and does not have distribution requirements for element durations, allowing us to incorporate information about the duration of sequence elements into our sequence analysis.

Like Markov chains and hidden Markov models, network models and formal grammars can also be used to assess and compare order, repetition and diversity of elements and they also neglect information about element duration (Kershenbaum et al., 2014). In contrast, temporal structure analysis focuses on timing information using a variety of techniques. Interpulse interval has been a widely used measure in temporal structure analysis, but it does not traditionally include order information. Multi-timescale rhythm analysis involves segmenting a recording and analysing each segment using the same signal-processing methods that measure acoustic features of individual notes (Saar & Mitra, 2008). In so doing, duration information of natural elements within the song can be lost.

Utility of Statistical Methodology to Analysis of Birdsong and Other Applications

The use of semi-Markov methods, Bayesian methods and classification trees to describe state transitions and durations does not have to be limited to the analysis of birdsong, or even to the analysis of communication structure. Indeed, the use of semi-Markov chains as a method of analysing general behaviour sequences has been described as early as 1959 (Cane, 1959). The use of Bayesian methods for estimating transition probabilities, in particular, is a reasonable approach (1) to prevent the assignment of zero probability to a transition that is possible but simply has not occurred in the current data set and (2) to prevent the assignment of one to a transition that is not necessarily inevitable. Classification trees can be used to ascertain not only whether differences exist between groups of sequences, but also to generate rules to categorize new sequences and to investigate how

sequence attributes are regulated. However, these statistical methods seem most appropriate to data that have a limited state space, large sample sizes and several replicates per sample.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Adret-Hausberger M, Jenkins PF. Complex organization of the warbling song in the European starling *Sturnus vulgaris*. Behaviour. 1988; 107(3):138–155.

Alger SJ, Maasch SN, Riters LV. Lesions to the medial preoptic nucleus affect immediate early gene immunolabeling in brain regions involved in song control and social behavior in male European starlings. European Journal of Neuroscience. 2009; 29(5):970–982. [PubMed: 19245371]

Alger SJ, Riters LV. Lesions to the medial preoptic nucleus differentially affect singing and nest box-directed behaviors within and outside of the breeding season in European starlings (*Sturnus vulgaris*). Behavioral Neuroscience. 2006; 120(6):1326–1336. [PubMed: 17201478]

Baker MC, Bjerke TK, Lampe H, Espmark Y. Sexual responses of female great tits to variation in size of males' song repertoires. American Naturalist. 1986; 128(4):491–498.

Balthazart J, Absil P, Gerard M, Appelants D, Ball GF. Appetitive and consummatory male sexual behavior in Japanese quail are differentially regulated by subregions of the preoptic medial nucleus. Journal of Neuroscience. 1998; 18(16):6512–6527. [PubMed: 9698339]

Balthazart J, Surlemont C. Copulatory behavior is controlled by the sexually dimorphic nucleus of the quail POA. Brain Research Bulletin. 1990; 25(1):7–14. [PubMed: 2207718]

Balthazart J, Surlemont C, Harada N. Aromatase as a cellular marker of testosterone action in the preoptic area. Physiology & Behavior. 1992; 51(2):395–409. [PubMed: 1557450]

Bensch S, Hasselquist D. Evidence for active female choice in a polygynous warbler. Animal Behaviour. 1992; 44(2):301–311.

Brenowitz EA, Margoliash D, Nordeen KW. An introduction to birdsong and the avian song system. Journal of Neurobiology. 1997; 33(5):495–500. [PubMed: 9369455]

Byers J, Hebets E, Podos J. Female mate choice based upon male motor performance. Animal Behaviour. 2010; 79:771–778.

Cane VR. Behaviour sequences as semi-Markov chains. Journal of the Royal Statistical Society B (Methodological). 1959; 21:36–58.

Catchpole CK. Temporal and sequential organization of song in sedge warbler (*Acrocephalus schoenobaenus*). Behaviour. 1976; 59:226–246. doi: 10.1163/156853976x00370.

Catchpole, CK.; Slater, PJB. Bird song: Biological themes and variations. 2. Cambridge University Press; Cambridge, U.K.: 2008.

Chaiken M. Call development in the European starling (*Sturnus vulgaris*): the structural and functional merging of two early call types. American Zoologist. 1983; 23(4):929.

Chaiken M, Gentner TQ, Hulse SH. Effects of social interaction on the development of starling song and the perception of these effects by conspecifics. Journal of Comparative Psychology. 1997; 111(4):379–392. [PubMed: 9419883]

De Groof G, Poirier C, George I, Hausberger M, Van der Linden A. Functional changes between seasons in the male songbird auditory forebrain. Frontiers in Behavioral Neuroscience. 2013; 7:196. doi:10.3389/fnbeh.2013.00196. [PubMed: 24391561]

Duffy DL, Bentley GE, Drazen DL, Ball GF. Effects of testosterone on cell-mediated and humoral immunity in non-breeding adult European starlings. Behavioral Ecology. 2000; 11(6):654–662.

Eens M. Understanding the complex song of the European starling: an integrated approach. Advances in the Study of Behavior. 1997; 26:355–434.

Eens M, Pinxten R, Verheyen RF. Temporal and sequential organization of song bouts in the starling. Ardea. 1988; 77:75–86.

Eens M, Pinxten R, Verheyen RF. Male song as a cue for mate choice in the European starling. Behaviour. 1991; 116:210–238.

Eens M, Pinxten R, Verheyen RF. Function of the song and song repertoire in the European starling (*Sturnus vulgaris*): an aviary experiment. Behaviour. 1993; 125:51–66.

Falk H, Gwinner E. Timing of photorefractoriness in the European starling: significance of photoperiod early and late in the reproductive cycle. Biology of Reproduction. 1988; 39(5):1004–1008. [PubMed: 3219372]

Franz M, Goller F. Respiratory patterns and oxygen consumption in singing zebra finches. Journal of Experimental Biology. 2003; 206(6):967–978. [PubMed: 12582139]

Gentner TQ, Hulse SH. Female European starling preference and choice for variation in conspecific male song. Animal Behaviour. 2000; 59:443–458. [PubMed: 10675267]

Hausberger, M. Social influences on song acquisition and sharing in the European starling (*Sturnus vulgaris*). In: Snowdon, CT.; Hausberger, M., editors. Social influences on vocal development. Cambridge University Press; Cambridge, U.K.: 1997. p. 128-156.

Henry L, Hausberger M, Jenkins PF. The use of song repertoire changes with pairing status in male European starling. Bioacoustics. 1994; 5(4):261–266.

Hill, T.; Pawel, L. A comprehensive reference for science, industry, and data mining. StatSoft; Tulsa, OK: 2006. Statistics: Methods and applications.

Jarrard LE. On the use of ibotenic acid to lesion selectively different components of the hippocampal formation. Journal of Neuroscience Methods. 1989; 29(3):251–259. doi: 10.1016/0165-0270(89)90149-0. [PubMed: 2477650]

Jin DZ, Kozhevnikov AA. A compact statistical model of the song syntax in Bengalese finch. PLoS Computational Biology. 2011; 7(3):e1001108. doi:10.1371/journal.pcbi.1001108. [PubMed: 21445230]

Jones AE, ten Cate C, Slater PJB. Early experience and plasticity of song in adult male zebra finches (*Taeniopygia guttata*). Journal of Comparative Psychology. 1996; 110(4):354–369.

Kershenbaum A. Entropy rate as a measure of animal vocal complexity. Bioacoustics. 2014; 23(3):195–208. doi:10.1080/09524622.2013.850040.

Kershenbaum A, Bowles AE, Freeberg TM, Jin DZ, Lameira AR, Bohn K. Animal vocal sequences: not the Markov chains we thought they were. Proceedings of the Royal Society B: Biological Sciences. 2014; 281(1792):20141370. doi:10.1098/rspb.2014.1370. [PubMed: 25143037]

Kvam, PH.; Vidakovic, B. Nonparametric statistics with applications to science and engineering. J. Wiley; Hoboken, NJ: 2007.

Lampe HM, Espmark YO. Singing activity and song pattern of the redwing *Turdus Iliacus* during the breeding season. Ornis Scandinavica. 1987; 18(3):179–185.

Martin DJ. Songs of the fox sparrow. 3. Ordering of song. Wilson Bulletin. 1990; 102(4):655–671.

Miksis-Olds JL, Buck JR, Noad MJ, Cato DH, Stokes MD. Information theory analysis of Australian humpback whale song. Journal of the Acoustical Society of America. 2008; 124(4):2385–2393. doi:10.1121/1.2967863. [PubMed: 19062876]

Mountjoy DJ, Lemon RE. Extended song learning in wild European starlings. Animal Behaviour. 1995; 49:357–366.

Nolan PM, Hill GE. Female choice for song characteristics in the house finch. Animal Behaviour. 2004; 67:403–410.

Nottebohm F, Nottebohm ME, Crane L. Developmental and seasonal changes in canary song and their relation to changes in the anatomy of song-control nuclei. Behavioral and Neural Biology. 1986; 46(3):445–471. [PubMed: 3814048]

Oberweger K, Goller F. The metabolic cost of birdsong production. Journal of Experimental Biology. 2001; 204(19):3379–3388. [PubMed: 11606611]

Riters LV, Ball GF. Lesions to the medial preoptic area affect singing in the male European starling (*Sturnus vulgaris*). Hormones and Behavior. 1999; 36(3):276–286. [PubMed: 10603291]

Riters LV, Eens M, Pinxten R, Duffy DL, Balthazart J, Ball GF. Seasonal changes in courtship song and the medial preoptic area in male European starlings (*Sturnus vulgaris*). Hormones and Behavior. 2000; 38(4):250–261. [PubMed: 11104643]

Rusch KM, Pytte CL, Ficken MS. Organization of agonistic vocalizations in black-chinned hummingbirds. Condor. 1996; 98(3):557–566. doi:10.2307/1369568.

Rusch KM, Thusius K, Ficken MS. The organization of agonistic vocalizations in ruby-throated hummingbirds with a comparison to blackchinned hummingbirds. Wilson Bulletin. 2001; 113(4): 425–430. doi:10.1676/0043-5643(2001)113[0425:Tooavi]2.0.Co;2.

Saar S, Mitra PP. A technique for characterizing the development of rhythms in bird song. PLoS One. 2008; 3(1):e1461. doi:10.1371/journal.pone.0001461. [PubMed: 18213370]

Sakata JT, Hampton CM, Brainard MS. Social modulation of sequence and syllable variability in adult birdsong. Journal of Neurophysiology. 2008; 99(4):1700–1711. doi:10.1152/jn.01296.2007. [PubMed: 18216221]

Scharff C, Nottebohm F. A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning. Journal of Neuroscience. 1991; 11(9):2896–2913. [PubMed: 1880555]

Searcy WA, Marler P. A test for responsiveness to song structure and programming in female sparrows. Science. 1981; 213(4510):926–928. [PubMed: 17775278]

Smith GT, Brenowitz EA, Beecher MD, Wingfield JC. Seasonal changes in testosterone, neural attributes of song control nuclei, and song structure in wild songbirds. Journal of Neuroscience. 1997; 17(15):6001–6010. [PubMed: 9221796]

Smith GT, Brenowitz EA, Wingfield JC. Seasonal changes in the size of the avian song control nucleus HVC defined by multiple histological markers. Journal of Comparative Neurology. 1997; 381(3): 253–261. [PubMed: 9133567]

ten Cate, C.; Lachlan, R.; Zuidema, W. Analyzing the structure of bird vocalizations and language: finding common ground. In: Bolhuis, JJ.; Everaert, M., editors. Birdsong, speech, and language: Exploring the evolution of mind and brain. MIT Press; Cambridge, MA: 2013. p. 243-260.

ten Cate C, Okanoya K. Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012; 367(1598):1984–1994. doi:10.1098/rstb.2012.0055.

van der Vaart, AW. Asymptotic statistics. Cambridge University Press; Cambridge, U.K.: 2000.

Van Hout AJ, Eens M, Balthazart J, Pinxten R. Complex modulation of singing behavior by testosterone in an open-ended learner, the European starling. Hormones and Behavior. 2009; 56(5): 564–573. doi:10.1016/j.yhbeh.2009.09.010. [PubMed: 19800345]

Verheyen, RF. Breeding strategies of the starling. In: Wright, EN., editor. Bird problems in agriculture. BCPC; London, U.K.: 1980. p. 69-82.

Watson JT, Adkins-Regan E. Testosterone implanted in the preoptic area of male Japanese quail must be aromatized to activate copulation. Hormones and Behavior. 1989; 23(3):432–447. [PubMed: 2793083]

Woolley SC, Doupe AJ. Social context-induced song variation affects female behavior and gene expression. PLoS Biology. 2008; 6(3):e62. doi:10.1371/journal.pbio.0060062. [PubMed: 18351801]

Highlights

- Vocal signals contain acoustic elements that differ in order and duration.

- We used novel statistical methods to examine this in male European starling songs.

- Songs of males in breeding and nonbreeding condition were statistically separable.

- Males in breeding condition produced sexually relevant attributes of song.

- Results also demonstrated a role for the medial preoptic nucleus in this process.
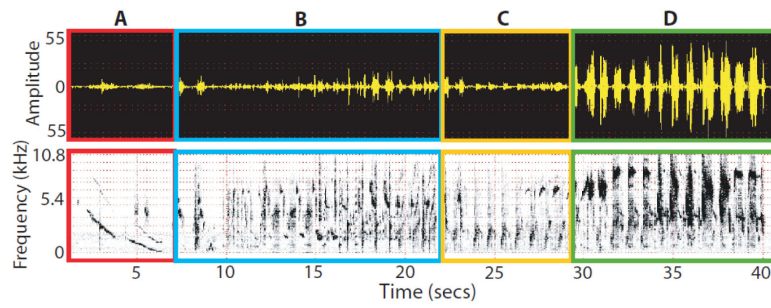
Figure 1.
Sample song oscillogram (top) and spectrogram (bottom) illustrating distinct categories of starling song consisting of (a) whistles, (b) complex phrases, (c) click phrases and (d) high-frequency trill phrases. See text for additional details.
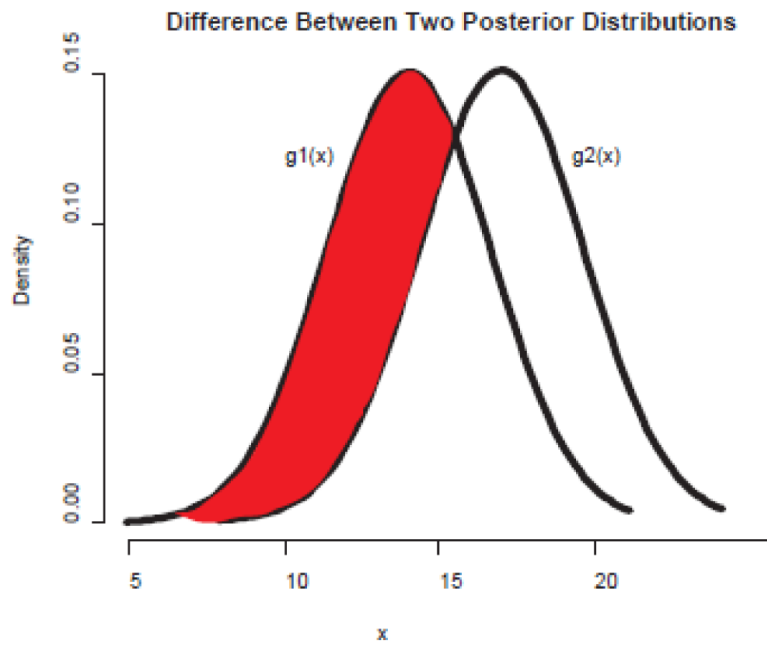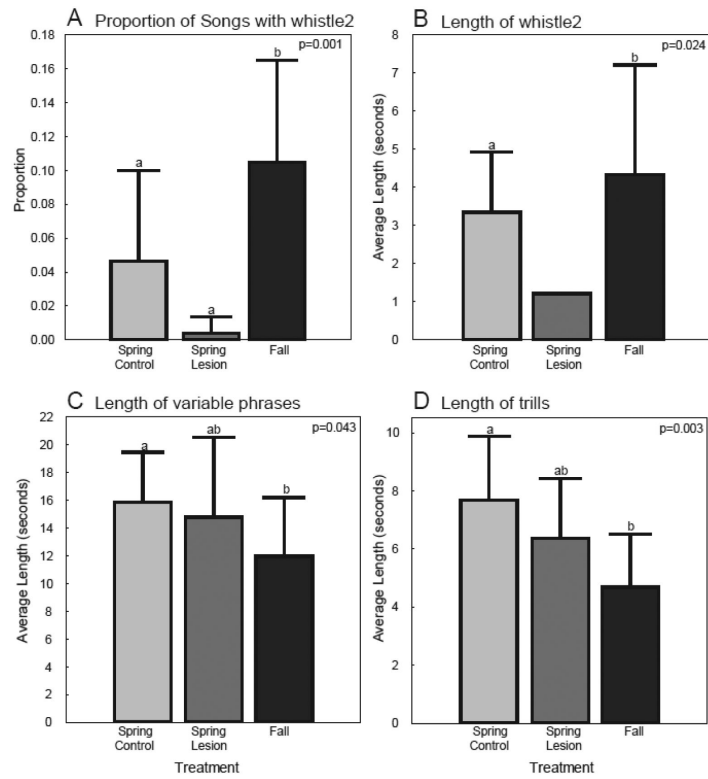
Figure 2.
An illustration of the difference between two hypothetical posterior distributions, $g_1(x)$ and $g_2(x)$. The shaded area under the $g_1(x)$ curve represents the difference between $g_1(x)$ and $g_2(x)$. Because both $g_1(x)$ and $g_2(x)$ are equal to 1, the difference between them is between 0 and 1.
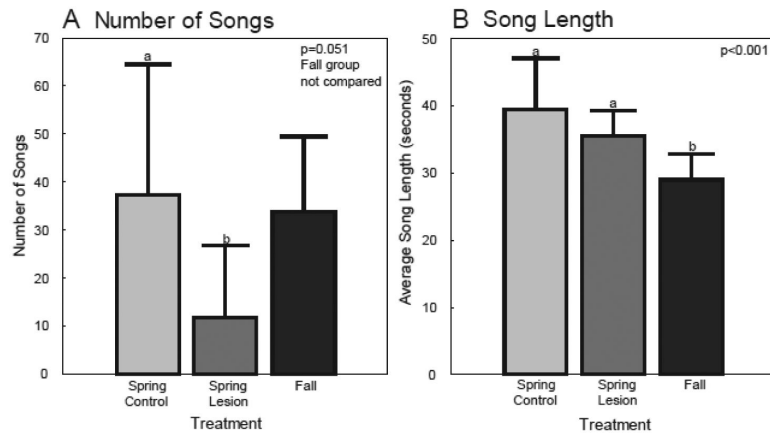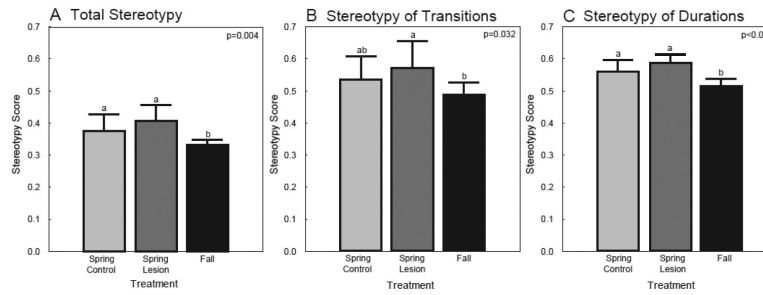
Figure 3.
Mean + SD (a) number of songs produced and (b) song bout length by bird treatment. Spring control and spring lesioned birds were all recorded for exactly 10 h. Autumn birds were recorded for variable periods of time and thus their number of songs was not compared statistically. *P* values are shown in the top right corner and significantly different post hoc results, where relevant, are indicated with different lowercase letters above the bars.

Figure 4.

Mean + SD (a) proportion of songs containing phrase category whistle2 per bird, (b) length of phrase category whistle2, (c) length of the variable phrase category and (d) length of trills. ANOVA *P* values are shown in the top right corner and significantly different post hoc results, where relevant, are indicated with different lowercase letters above the bars.

Figure 5.
Mean + SD (a) total stereotypy score per bird, (b) transition stereotypy score per bird and (c) duration stereotypy score per bird. ANOVA *P* values are shown in the top right corner and significant post hoc results are indicated with lowercase letters above the bars.

Table 1

State transition structure

k		W1	W2	WS	VP	C	T	Stop
5	Start		0					0
3	Whistle1 (W1)	0	0	0				0
1	Whistle2 (W2)	0	0	0	0	0	0	1
1	WhistleSong (WS)	0	0	0	0	0	0	1
5	Variable phrase (VP)			0	0			
5	Clicks (C)			0		0		
5	Trills (T)			0			0	

Table 2

Classification tree rules for distinguishing spring control from autumn male song

Song structure variable	Deviance	Rules that indicate spring song	Spring control birds that followed rule	Spring lesioned birds that followed rule	Autumn birds that followed rule
Mean duration of trills	17.51	4.98 s	11/12	4/6	2/12
Mean duration of variable phrases	19.87	10.05 s	11/12	5/6	5/12
Mean duration of clicks	22.91	8.41 s	5/12	2/6	0/12
Standard deviation of the duration of clicks	22.91	<1.75 s	5/12	0/6	0/12
Probability of transitioning from trills to whistle1	25.01	<0.003	4/12	1/6	0/12
Probability of transitioning from trills to whistle2	25.01	<0.003	4/12	0/6	0/12
Probability of transitioning from clicks to whistle2	25.12	<0.051	10/12	6/6	5/12

Variables are listed in the order of strength of differentiating spring control from autumn male song, and where log durations were used in analysis, durations are reported. The proportion of birds that followed the rules in each treatment group is listed.