JOURNAL OF AVIAN BIOLOGY

Song but not plumage varies geographically among Willow Flycatcher (Empidonax traillii) subspecies

Journal:	*Journal of Avian Biology*
Manuscript ID	JAV-02621.R1
Wiley - Manuscript type:	Research Article
Keywords:	song variation, plumage variation, subspecies, Willow Flycatcher
Abstract:	Plumage and song are important signals used by birds to attract mates and repel rivals. Divergence in sexual signals can lead to reproductive isolation among incipient species, but the relative importance of each modality may vary among taxa. Tyrannid flycatchers exhibit evolutionarily conservative plumage coloration but distinct song structure among subfamilies and species. Thus, tyrannids are an interesting group in which to study the relative role of plumage and song in contributing to population divergence. In this study, we assessed character divergence among four Willow Flycatcher (Empidonax traillii) subspecies by measuring spectral reflectance of plumage modeled in tetrahedral colorspace from museum specimens collected on putative breeding grounds. We also quantified differences in song structure based on publicly available and field-recorded songs across the species range. Using unsupervised and unbiased clustering algorithms that assigned group membership independent of a priori taxonomic designations, we found that currently recognized subspecies did not consistently sort in accordance with subspecies designation based on plumage color. However, song analyses grouped birds into two clusters; one that included 89% of all putative E. t. extimus, and another that included 100% of specimens designated as E. t. adastus, E. t. brewsteri, E. t. traillii, and a small percentage of E. t. extimus (11%). Our results are consistent with previous hypotheses of conservative plumage evolution in tyrannids and species differentiation based on song, and supports the subspecific status of E. t. extimus.

JOURNAL OF AVIAN BIOLOGY

Article

Song but not plumage varies geographically among willow flycatcher *Empidonax traillii* subspecies

Sean M. Mahoney, Matthew W. Reudink, Bret Pasch and Tad C. Theimer

S. M. Mahoney (https://orcid.org/0000-0001-8446-423X) ✉ (sean.mahoney@nau.edu), B. Pasch and T. C. Theimer, Dept of Biological Sciences, Northern Arizona Univ., Flagstaff, AZ, USA. – M. W. Reudink, Dept of Biological Sciences, Thompson Rivers Univ., Kamloops, BC, Canada.

Journal of Avian Biology
2020: e02621
doi: 10.1111/jav.02621

Subject Editor: Martin Paeckert
Editor-in-Chief: Thomas Alerstam
Accepted 29 October 2020

Plumage and song are important signals used by birds to attract mates and repel rivals. Divergence in sexual signals can lead to reproductive isolation among incipient species, but the relative importance of each modality may vary among taxa. Tyrannid flycatchers exhibit evolutionarily conservative plumage coloration but distinct song structure among subfamilies and species. Thus, tyrannids are an interesting group in which to study the relative role of plumage and song in contributing to population divergence. In this study, we assessed character divergence among four willow flycatcher *Empidonax traillii* subspecies by measuring spectral reflectance of plumage modeled in tetrahedral colorspace from museum specimens collected on putative breeding grounds. We also quantified differences in song structure based on publicly available and field-recorded songs across the species range. Using unsupervised and unbiased clustering algorithms that assigned group membership independent of a priori taxonomic designations, we found that currently recognized subspecies did not consistently sort in accordance with subspecies designation based on plumage color. However, song analyses grouped birds into two clusters; one that included 89% of all putative *E. t. extimus*, and another that included 100% of specimens designated as *E. t. adastus*, *E. t. brewsteri*, *E. t. traillii* and a small percentage of *E. t. extimus* (11%). Our results are consistent with previous hypotheses of conservative plumage evolution in tyrannids and species differentiation based on song, and supports the subspecific status of *E. t. extimus*.

Keywords: plumage variation, song variation, subspecies, willow flycatcher

Introduction

Many animals use acoustic and visual signals to discriminate between conspecifics and heterospecifics in various social interactions (Bradbury and Vehrencamp 1998). Signal divergence among populations resulting from natural selection, sexual selection, genetic drift, cultural drift, environmental conditions (Hill 2006, Wilkins et al. 2013), arbitrary mate choice (Prum 2010) or some combination of all these phenomena can be important in reproductive isolation and speciation (Mayr 1942). Although genomic techniques are often used to delimit species (Barrowclough et al. 2016), acoustic and visual characters can be a powerful alternative (Mallet 2005, Tobias et al. 2010). In birds, plumage differences are often used to delimit species (e.g. *Dendroica* warblers,

NORDIC SOCIETY OIKOS
www.avianbiology.org

Shutler and Weatherhead 1990, Lovette and Bermingham 1999, Toews et al. 2016) but when plumage variation is limited, song may provide a more robust indicator of species limits (Martens et al. 2003, Päckert et al. 2003, Rheindt et al. 2008).

Plumage evolution in tyrannid flycatchers is evolutionarily conservative (Zink and Johnson 1984) resulting in a family of many sibling species that are difficult to differentiate by plumage alone (Stein 1958, 1963, Rheindt et al. 2008). Morphologically similar species are differentiated instead by differences in song or behavior (Stein 1963, Johnson and Cicero 2002, Rheindt et al. 2008). In spite of the general lack of strong plumage variation among *Empidonax* flycatchers, subspecies within this genus were often discriminated by plumage. For example, the original subspecies designations of the four subspecies of the willow flycatcher currently recognized by the US Fish and Wildlife Service (USFWS; *E. t. traillii* in the eastern US, *E. t. brewsteri* in the northwestern US, *E. t. adastus* in the interior west and *E. t. extimus* in southwest riparian areas) were based on qualitative plumage descriptions (Brewster 1895, Oberholser 1918, 1932, 1947, Phillips 1948, Aldrich 1951, Unitt 1987, Browning 1993). These studies were subsequently criticized due to a lack of rigorous statistical analyses, small sample sizes, and conclusions based on specimens collected over a relatively small geographic area (Zink 2015). More recently, a study using a colorimeter to measure plumage reflectance found differences among subspecies in mean values on the back and crown but substantial overlap among individuals (Paxton et al. 2010). The colorimeter used in that study did not assess UV reflectance, which is an important part of visual signals in birds generally (Cuthill et al. 2000), and may contribute to sexual dichromatism in willow flycatchers specifically (Eaton 2007). Paxton et al. (2010) was subsequently criticized because measurements were based on wild birds captured and released in the field and therefore specimens were unavailable for later reanalysis (Zink 2015). Only one study has rigorously assessed geographic song variation in willow flycatchers; Sedgwick (2001) found differences in song structure between *E. t. adastus* and *E. t. extimus* but only compared two subspecies and songs were not catalogued nor made publicly available. Finally, all studies assessing willow flycatcher phenotypic variation based analyses on specimens grouped a priori into presumed subspecies and so comparisons were not made independent of taxonomic identity. The validity of subspecific designations in this group have important implications because one subspecies *E. t. extimus* is listed as endangered (US Fish and Wildlife Service, USFWS 1995) and uncertainty around subspecies designation hinders continued conservation protection (Haig et al. 2006, Zink 2015, Theimer et al. 2016).

In this study, we quantitatively compare plumage variation across all four subspecies of willow flycatchers using museum specimens and methods that include reflectance into the UV spectrum and incorporate models of avian visual sensitivity. These methods account for the potential presence of plumage signals that would not be detected by human-biased assessments of color based on features like lightness, saturation and hue used in previous analyses (Paxton et al. 2010). Investigation of the potential for UV signal divergence is particularly warranted given that Eaton (2007) found evidence for sexual dichromatism in UV signals from both the back and belly of *Empidonax traillii* (Eaton 2007, their Supporting information). In addition, we expand on Sedgwick's (2001) original comparison of songs of *E. t. extimus* and *E. t. adastus* by including songs of all four subspecies using publicly available recordings. Importantly, we utilize unbiased and unsupervised analyses that assess differences independent of taxonomic identity.

Methods

Reflectance measurements

To assess plumage differences among willow flycatcher subspecies, we measured plumage reflectance across the avian visual range (300–700 nm) on the back, belly, breast, crown, nape and throat of willow flycatchers (*E. t. adastus*, n = 38; *E. t. brewsteri*, n = 30; *E. t. extimus*, n = 21; *E. t. traillii*, n = 18) and, as an outgroup, alder flycatcher (n = 8) museum specimens (Supporting information). Specimens were collected from across the breeding range of Willow and alder flycatchers (Fig. 1A). We restricted the specimens included in our study to adults collected during the peak breeding season (15 June–25 July). We assigned each specimen to a subspecies based on its collection location using the putative subspecies maps from USFWS (1995), Paxton (2000) and Paxton et al. (2008). Specimens were not genotyped. We measured reflectance using a JAZ spectrometer (Ocean Optics, Dunedin, FL) with a fiber optic reflectance probe and xenon pulsed light source. To minimize detection of ambient light, the probe was housed in a sheath that held the probe at a 90° angle and at a fixed distance of 5.9 mm from the feather surface. Specimens were placed flat on a low reflectance matte black sheet of paper (e.g. construction paper) and we then took 10 replicate readings throughout the area of each plumage patch (Reudink et al. 2009). Reflectance spectra were recorded using SpectraSuite (ver. 1.0, Ocean Optics). We standardized the reflectance measures between each patch by measuring a white (Ocean Optics WS-1) and dark (sealed, black velvet-lined box) standard. Prior to reflectance measurements, reference photos of each specimen were taken (Fig. 2B).

Song recordings

To assess geographic song variation of willow flycatcher subspecies, we used song recordings from across the US willow flycatcher range (Fig. 2A, *E. t. adastus* n = 46; *E. t. brewsteri* n = 32; *E. t. extimus* n = 37; *E. t. traillii* n = 26). We acquired publicly available songs from the Cornell Lab of Ornithology Macaulay Library and xeno-canto.org libraries (Supporting information). We restricted our recordings to the 'fitz-bew' vocalization, as that is used to attract mates and defend

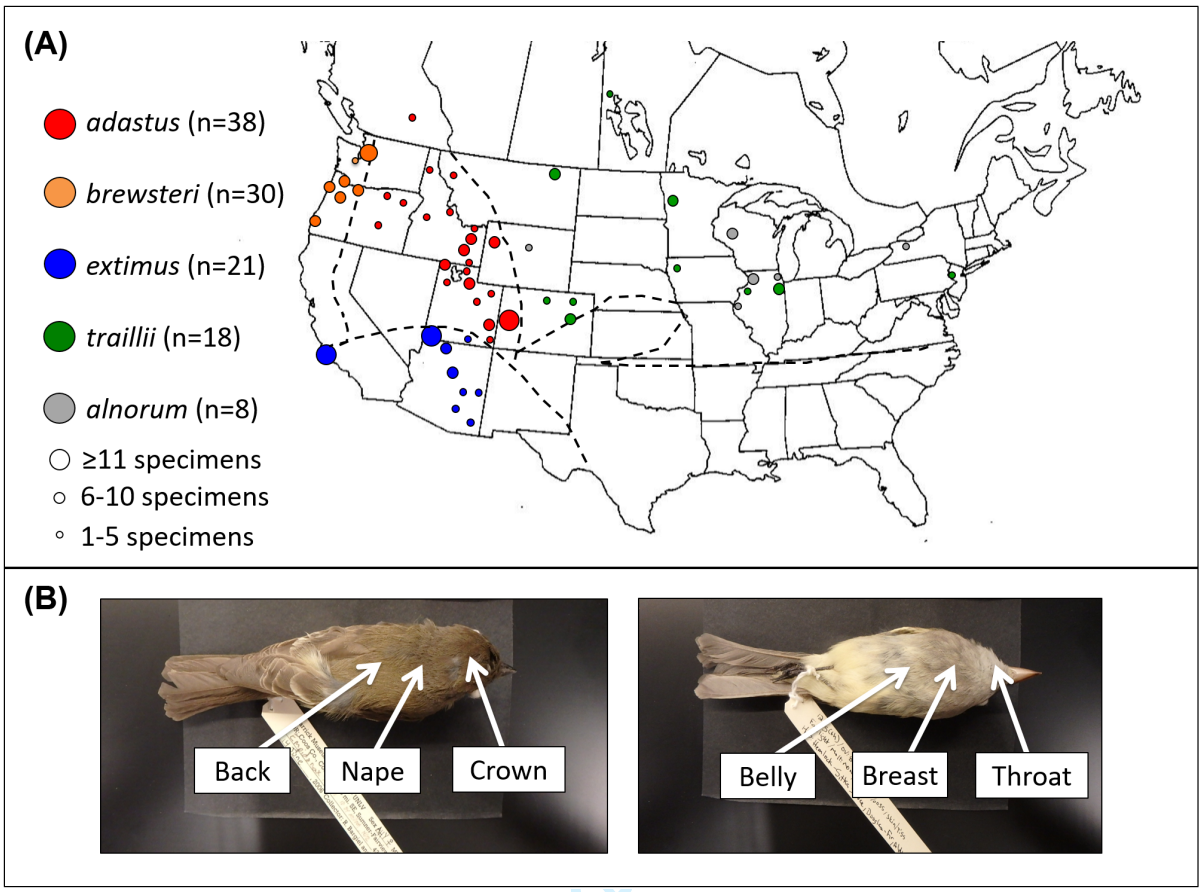

Figure 1. (A) Specimen collection locations of willow flycatcher subspecies (*Empidonax traillii adastus*, red (n = 38), *E. t. brewsteri*, orange (n = 30), *E. t. extimus*, blue (n = 21), *E. t. traillii*, green (n = 18)). Subspecies were assigned a priori based on collection location between 15 June and 25 July. Dots are scaled to sample size. Dashed lines indicate putative subspecies boundaries. (B) Locations of dorsal (left) and ventral (right) plumage patches of museum specimens on which light reflectance (300–700 nm) was measured to assess subspecific plumage differences.

territories (Stein 1958, Prescott 1987). To avoid recordings from non-breeding migrants, only songs that were recorded during June and July and that included location metadata were included in analyses. We included n = 6 songs in late May (*E. t. adastus* n = 1, *E. t. extimus* n = 2, *E. t. traillii* n = 3), but only when the recorder noted in the metadata that the individual bird was actively breeding (based on observing either a mated pair and/or an active nest). We also recorded singing male willow flycatchers on their breeding territories (Supporting information, locations referenced from Paxton 2000 and Paxton et al. 2008) during June and July 2016–2019 using a Sennheisser ME66 shotgun microphone (mono-line) with Rycote handgrip and Rycote Softie windshield and Marantz PMD661 MKII solid-state recorder. Our sampling rate (frequency samples per second) was set to 44.1 kHz.

Quantifying song characteristics

We digitized willow flycatcher songs using Raven Pro (Cornell Lab of Ornithology) using the Hann window (size = 256 samples, 50% window overlap, DFT = 256 samples) with unsmoothed view (Fig. 2B). We then quantified 44 acoustic parameters of the 'fitz-bew' vocalization (Supporting information): 26 measures of frequency, 16 measures of duration and two measures of frequency modulations (adapted from Sedgwick 2001). To minimize background noise, all songs were high- and low-pass filtered at 1 kHz and 7 kHz respectively. Only songs with high signal-noise ratio were used in analyses and we did not include recordings that were missing a parameter or in which the quality was too low to measure a parameter.

Data analysis

Tetrahedral colorspace analysis

We analyzed raw spectral data using the *pavo* package (Maia et al. 2013) for *R* (<www.r-project.org>). We corrected negative values by adding the minimum reflectance value to all spectra (Maia et al. 2013). To assess plumage variation among willow flycatcher subspecies and alder flycatchers, we modeled the reflectance data in 'tetrahedral colorspace' using avian visual models that correct for differences in avian visual systems (Vorobeyev et al. 1998). Although

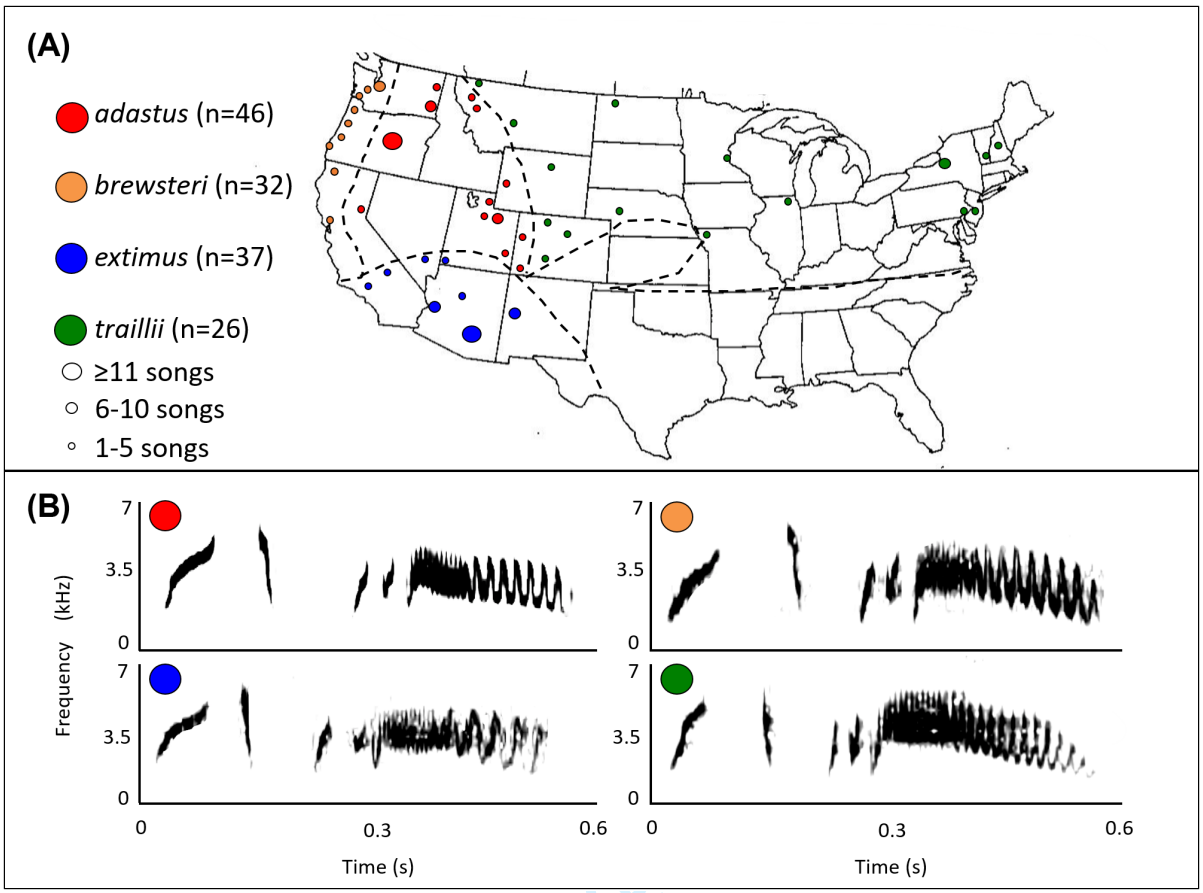

Figure 2. (A) Recording locations and sample sizes of willow flycatcher subspecies (*Empidonax traillii adastus*, red (n = 46), *E. t. brewsteri*, orange (n = 32), *E. t. extimus*, blue (n = 37), *E. t. traillii*, green (n = 26)). Subspecies were assigned a priori based on recording location between 15 June and 25 July. Dots are scaled to sample size. Dashed lines indicate putative subspecies boundaries. (B) Representative spectrograms of willow flycatcher subspecies song. Colored dots indicate subspecies from panel A. Representative songs were selected based on songs that fell in the approximate center of the subspecies clustering in PCA space.

the visual sensitivity range of willow and alder flycatchers is unknown, visual systems of other Tyrannidae is biased towards violet wavelengths (Ödeen and Håstad 2003), so our tetrahedral models estimated coordinates that correspond to the average violet sensitive avian photoreceptor sensitivity to violet (v), short (s), medium (m) and long (l) wavelengths. We then summarized the tetrahedral data using a non-metric multidimensional scaling (NMDS) analysis. To test for sexual dichromatism (Eaton 2007) in tetrahedral space and the effects of specimen age, we used separate mixed effects models. NMDS1 was our response variable and taxa, sex (female: n = 41; male: n = 66), specimen age, and the interaction between taxa and sex were included as fixed effects. Museum was included as a random effect. We treated plumage patches as distinct body regions and therefore separate hypotheses so we did not adjust our alpha level for multiple comparisons (Montgomerie 2006). We excluded specimens from this analysis where sex was unknown (n = 8).

Song structure analysis

Because many of our acoustic parameters were collinear, we summarized song characteristics with a principal components analysis (PCA), using the *prcomp* command in *R* (<www.r-project.org>). Prior to PCA we tested raw data for homoscedasticity using a Levene's test ($F_{3,137} = 0.33$, $p = 0.80$) and to meet assumptions of linearity, we log transformed all acoustic parameters. To help with interpretation of the loadings, we multiplied −1 to PC scores (Vehrencamp et al. 2003).

Clustering models, and group assignment analyses

To assess whether taxa occupied different tetrahedral color space or produced unique songs, we used an unbiased and unsupervised classification approach that is independent of taxa identification. First, we tested the null hypothesis that there are no groupings of plumage color and song characteristics – and therefore the data fit best within one group cluster – using the *fvis_nbclust* function in the *factoextra* package for *R* (Kassambara and Mundt 2017). This analysis assesses the quality of group clusters (i.e. how well the data fit within clusters) by calculating the silhouette width for n = 1–5

clusters for plumage color and n = 1–4 clusters for song. The silhouette width is a measure of confidence for group membership within a cluster and values range from −1 to +1 with values closer to 1 represent better clustering (Rousseeuw 1987). For all plumage patches and song, groupings were optimized in n > 1 clusters (Supporting information), suggesting some clustering of plumage color and song characteristics. Next, we determined the appropriate number of group clusters of NMDS color space scores and song PCA scores using *ClValid* (Brock et al. 2008) which evaluates clustering models and the numbers of clustering groups independent of taxa identity and subsequently identifies the appropriate clustering algorithm. In our analyses, we evaluated K-means and hierarchical clustering models with n = 2–5 clustering groups for plumage color and n = 2–4 for song. *ClValid* assesses group clustering based on three indices: connectivity, Dunn and silhouette width. The connectivity index assigns group membership of data points based on the proximity to other samples. Connectivity ranges from 0 to infinity and smaller values represent well clustered data (Handl et al. 2005). The Dunn and silhouette indices are measures of the 'compactness' and 'spread' of clusters. The Dunn metric is the ratio between the smallest distance between data points from different clusters and the largest intracluster distance (Dunn 1974). Dunn indices range from 0 and infinity and higher values represent better clustering. Silhouette values estimate the degree of confidence in membership within a particular cluster (Rousseeuw 1987). The silhouette indices are estimated by calculating the mean distance of points within a cluster and the mean distance between points and range from −1 to +1 and values close to 1 represent better clustering. Therefore, we chose the number of groups and the clustering method in our analyses based on models with optimized connectivity, Dunn and silhouette values (Brock et al. 2008). We selected the best clustering algorithm based on a rank-based optimization approach using the *RankAggreg* package in *R* (Pihur et al. 2009). *RankAggreg* generates a rank-based list corresponding to the performance of a given measure and then provides an overall ranking from a Spearman's footrule analysis, which minimizes distance from all constituent lists (Bible et al. 2013). Similar clustering analyses have been used in previous studies with acoustic (Jeon and Hong 2015) and morphological data (Lucek et al. 2016). Based on the clustering method results, we used either hierarchical clustering using the *ward.D2* function in the *hclust* package or K-means clustering using the *kmeans* function in R to evaluate agreement between clustering group assignment and taxa identity. To further assess plumage color variation among taxa, we plotted individual specimens in tetrahedral colorspace which models specimen color relationships and corrects for avian visual systems.

Results

Plumage

For all plumage patches, our clustering selection analyses identified two clustering groups ('plumage color group 1 and 2') and hierarchical clustering as the best algorithm (Table 1), so our classifications are based on those models.

Although the clustering methods identified two groups based on NMDS scores (Table 1, Supporting information) for all plumage patches, we found low agreement between taxonomic identity (based on geographic region where the specimen was collected) and colorspace grouping (Fig. 3). On all patches, taxa overlapped in tetrahedral colorspace (Fig. 4).

We found no evidence for sexual dichromatism on the back (Supporting information, $F_{1,93.3} = 3.5$, $p = 0.06$), belly

Table 1. Results from clustering algorithm selection for willow *Empidonax traillii* and alder flycatcher *Empidonax alnorum* plumage and willow flycatcher song. Clustering group number and algorithm was selected based on NMDS tetrahedral colorspace scores and song PCA scores using *ClValid* and *AggRankreg* which evaluates clustering models and the numbers of clustering groups independent of taxa identification. In our analyses, we evaluated K-means and hierarchical clustering models with n = 2–5 (plumage) and n = 2–4 (song) clustering groups.

	Index[1]	Score	Method[2]	Cluster
Back	Connectivity	4	Hierarchical	2
	Dunn	0.18	Hierarchical	2
	Silhouette	0.43	Hierarchical	2
Belly	Connectivity	4	Hierarchical	2
	Dunn	0.18	Hierarchical	2
	Silhouette	0.43	Hierarchical	2
Breast[2]	Connectivity	5.8	Hierarchical	2
	Dunn	0.15	Hierarchical	3
	Silhouette	0.6	Hierarchical	2
Crown	Connectivity	6.13	Hierarchical	2
	Dunn	0.15	Hierarchical	2
	Silhouette	0.62	Hierarchical	2
Nape[2]	Connectivity	4.2	Hierarchical	2
	Dunn	0.19	Hierarchical	5
	Silhouette	0.58	Hierarchical	2
Throat	Connectivity	2.9	Hierarchical	2
	Dunn	0.4	Hierarchical	2
	Silhouette	0.8	Hierarchical	2
Song	Connectivity	2.92	Hierarchical	2
	Dunn	0.65	Hierarchical	2
	Silhouette	0.59	Hierarchical	2

[1] The connectivity index assigns group membership of data points based on the proximity to other samples, where relatively smaller connectivity metrics indicate well-clustered groups. The Dunn metric is an estimate of the intercluster distance relative to intracluster distance of data points, where values > 1 indicate well-clustered data. And the silhouette index is the mean distance of points within a cluster and the mean distance between points, where silhouette values > 1 indicates well-clustered data.

[2] Clustering algorithm and cluster number was selected based on overall rank-based selection process using the *RankAggreg* package in *R* (Pihur et al. 2009). For the breast and nape, *RankAggreg* identified Hierarchical clustering with two groups (Spearman distances breast: 6.07; nape: 7.50).

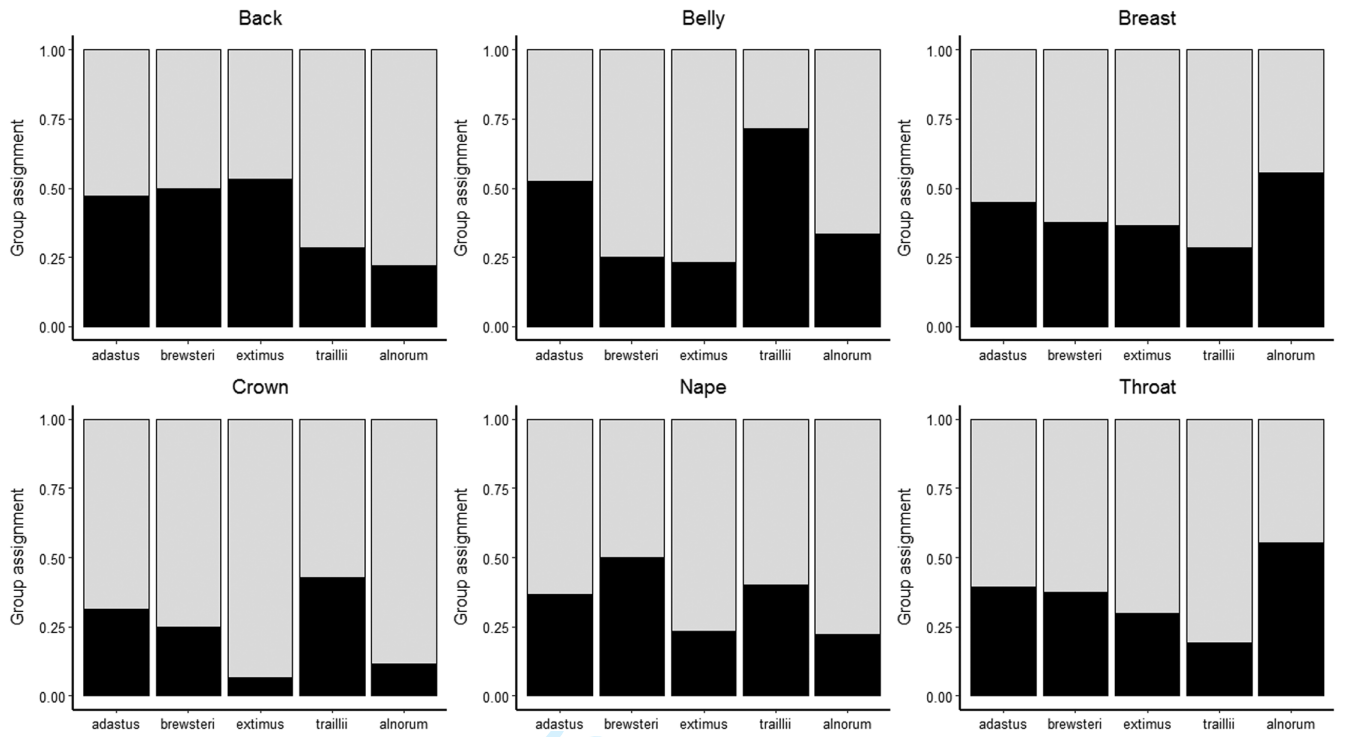

Figure 3. Proportion of plumage color group assignments for the back, belly, breast, crown, nape and throat (black, 'plumage color group 1', gray, 'plumage color group 2') of willow flycatcher subspecies (*Empidonax traillii adastus*, *E. t. brewsteri*, *E. t. extimus*, *E. t. traillii*) and alder flycatcher *E. alnorum* museum specimens. Group assignments were based on unsupervised and unbiased classification of non-metric multidimensional scaling (NMDS) of tetrahedral colorspace modeling that corrects for avian visual systems. Among all taxa, group assignment varied by individual.

(Supporting information, $F_{1,91.2}=0.07$, $p=0.79$), breast (Supporting information, $F_{1,93.0}=1.66$, $p=0.2$), nape (Supporting information, $F_{1,92.5}=0.05$, $p=0.83$), however on the crown, males averaged higher colorspace scores than females (Fig. 5, Supporting information, $F_{1,94.5}=8.02$, $p=0.006$). There was an interaction between taxa and sex on the throat (Supporting information, $F_{4,91.2}=2.5$, $p=0.04$), meaning the direction of dichromatism differed among taxa. Whereas *E. t. adastus* and *E. alnorum* males exhibited lower average throat colorspace scores than females, *E. t. brewsteri*, *extimus* and *traillii* males averaged higher throat colorspace scores than females (Fig. 5). There was a significant effect of specimen age on the breast (Supporting information, $F_{1,95.9}=6.59$, $p=0.01$) and throat (Supporting information, $F_{1,54.6}=14.6$, $p=0.0003$). Museum explained on average relatively little variation for all patches (Supporting information, mean $\sigma^2=0.0009$, range $\sigma^2=0.0001$–0.004).

Song

From our principal components analysis, PC1 explained 38% of the variation. Positive PC1 scores were associated with songs exhibiting higher overall minimum frequencies and higher frequencies in Phrase 1 note 2 and Phrase 2 note 1 (Supporting information). Negative PC1 scores were associated with songs with lower overall minimum frequencies and lower frequencies in Phrase 1 note 2 and Phrase 2 note 1 (Supporting information). PC2 explained 12% of the variation and positive scores were associated with songs containing more frequency modulations in the terminal portion of the song and higher frequencies in Phrase 1 note 1. Negative PC2 scores were associated with songs containing fewer terminal frequency modulations and lower frequencies in Phrase 1 note 1 (Supporting information).

Our song clustering analysis identified two groups ('song group 1 and 2') and hierarchical clustering as the best algorithm (Table 1) so our classifications are based on that model. When we compared group membership to taxonomic identification based on geographic origin of each song, we found song group 1 contained exclusively *E. t. extimus* songs (n = 33, 89%; Fig. 6). Song group 2 contained high proportions of *E. t. adastus* (n = 46, 100%; Fig. 6), *E. t. brewsteri* (n = 32, 100%; Fig. 6) and *E. t. traillii* (n = 26, 100%; Fig. 6) and relatively low proportions of *E. t. extimus* (n = 4, 11%; Fig. 6). Song groups separated in song PC space (Fig. 7).

Discussion

Using tetrahedral colorspace models, we found plumage color did not differ among willow flycatcher subspecies or between willow and alder flycatchers. Interestingly, we found evidence for sexual dichromatism on the crown and throat but not on the back or belly as was previously proposed by Eaton

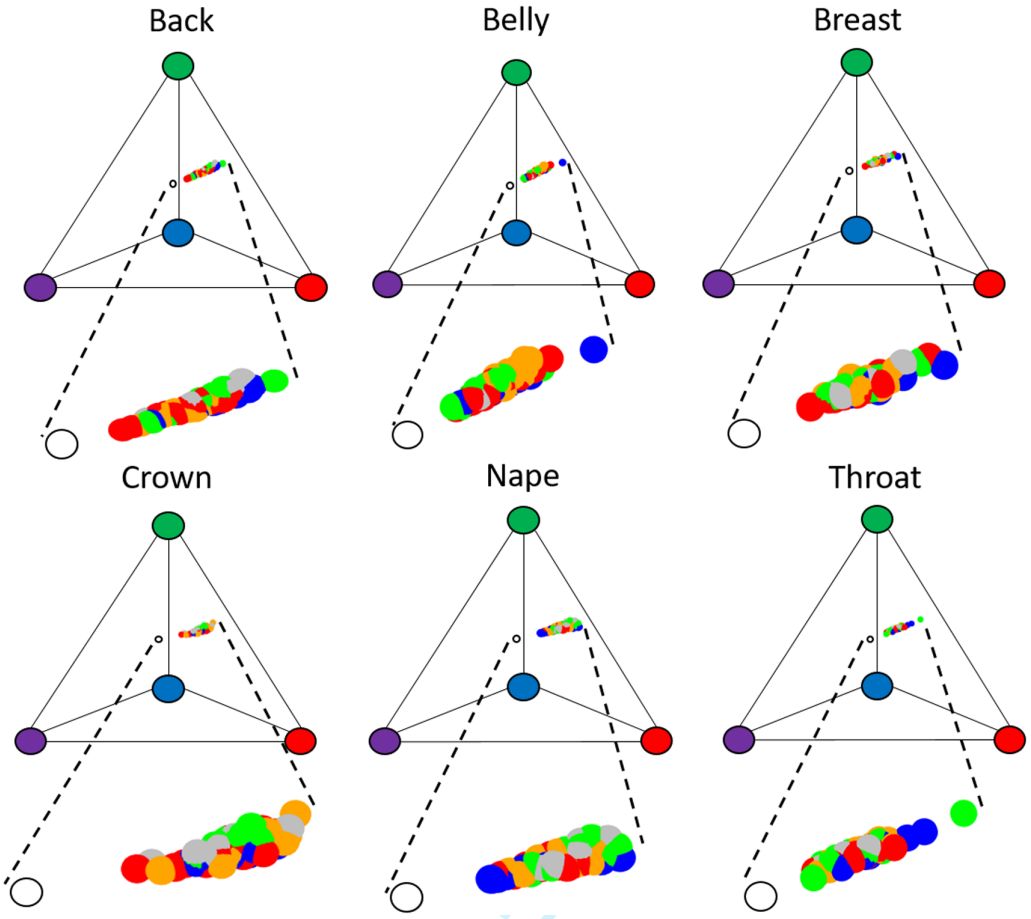

Figure 4. Plumage variation among willow flycatcher *Empidonax traillii* subpsecies (*E. t. adastus*, red (n = 38), *E. t. brewsteri*, orange (n = 30), *E. t. extimus*, blue (n = 21), *E. t. traillii*, green (n = 18)) and alder flycatchers (*E. alnorum*, gray (n = 8)) as modeled in 3D tetrahedral colorspace relative to an achromatic center (white dot) for back, belly, breast, crown, nape and throat plumage patches. For all plumage patches, taxa overlap in colorspace.

(2007), suggesting that these plumage patches may be important in social interactions (Hunt et al. 1999, Griffith et al. 2003, Alonso-Alvarez et al. 2004, Limbourg et al. 2004). In general, males averaged higher crown colorspace scores than females, but the direction of throat dichromatism differed in *E. t adastus* and *E. alnorum* relative to the other flycatcher taxa. Our plumage results differ from previous research that found color differences based on both qualitative (Phillips 1948, Aldrich 1951, Unitt 1987) and quantitative methods (Paxton et al. 2010). In part these differences may have been due to the fact that we used tetrahedral color models corrected for avian visual systems and included UV reflectance, whereas previous studies focused on differences based on human visual systems. One caveat of our plumage analysis was that we used museum study skins rather than live birds, and plumage spectra can differ between wild and museum specimens (McNett and Marchetti 2005, Doucet and Hill 2009). Time of year that specimens were collected (Doucet and Hill 2009), chemicals used to preserve skins (Pohland and Mullen 2006), specimen age (Armenta et al. 2008, Doucet and Hill 2009) and anthropogenic landscape conversions (Mason and Unitt 2018) have all been associated with changes in reflectance in museum skins. After correcting for avian visual systems, we found color differences among flycatcher taxa on the back, belly and throat. However, there was a significant effect of specimen age on throat color, so caution should be observed when interpreting those results. Although a study of several species of wood warblers found fading was minimal for specimens collected in the past 50 years (Armenta et al. 2008) only 14 of the 115 specimens we examined were less than 50 years of age. Therefore, the potential exists that live birds may show differences among subspecies that we failed to detect in museum skins, although the considerable overlap in reflectance values documented in the one study that did examine reflectance from live flycatchers (Paxton et al. 2010) suggests the unsupervised and unbiased clustering algorithm approach we used would have failed to detect strong groupings in data from live birds.

In contrast to plumage, song structure varied among flycatcher subspecies with *E. t. extimus* singing songs different in structure relative to the other willow flycatcher subspecies. Although the majority of putative *E. t. extimus* grouped into

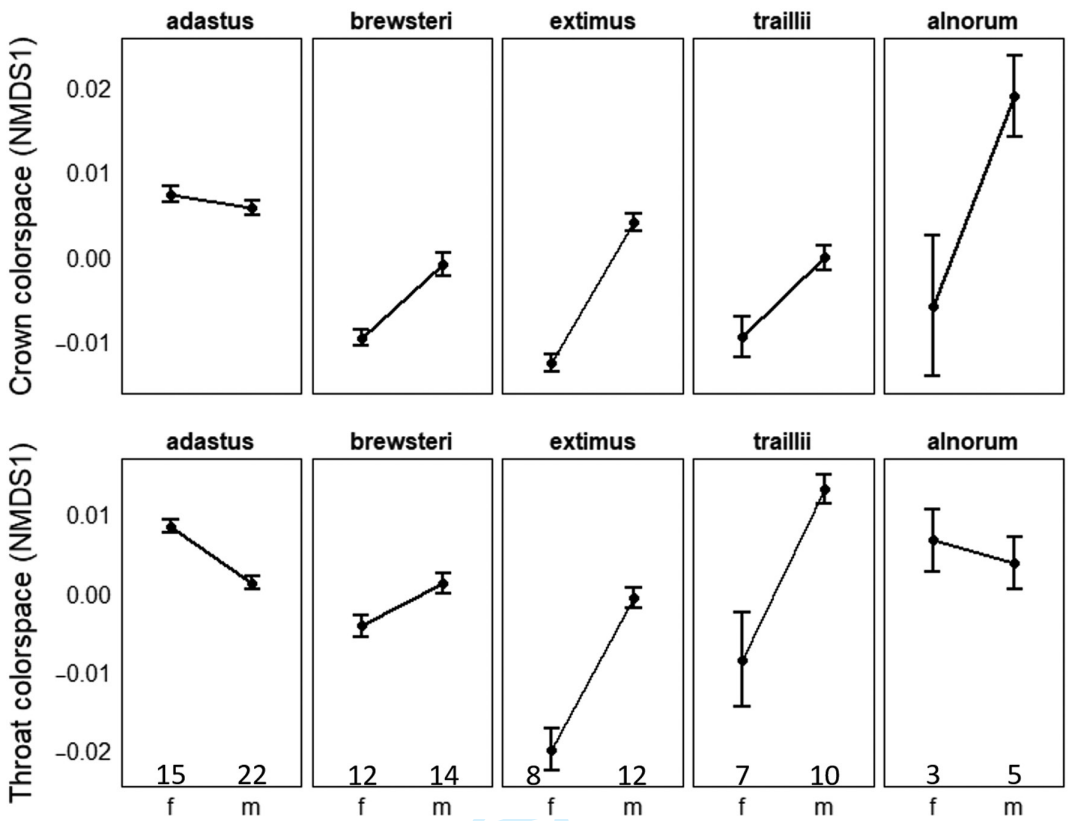

Figure 5. Mean (± SE) NMDS1 scores representing crown colorspace (top panel) and throat colorspace (bottom panel) for female (f) and male (m) willow flycatcher *Empidonax traillii* subspecies (*E. t. adastus*, *E. t. brewsteri*, *E. t. extimus*, *E. t. traillii*) and alder flycatchers *Empidonax alnorum*. There was a significant effect of sex on the crown ($F_{1, 94.5} = 8.02$, $p = 0.006$) and an interaction between sex and taxa on the throat ($F_{4, 91.2} = 2.5$, $p = 0.04$). NMDS1 colorspace scores were calculated from tetrahedral color models which model plumage color after correcting for avian visual systems. Sample sizes are indicated in bottom panel on x-axis. n = 8 specimens were excluded from this analysis because sex was unknown.

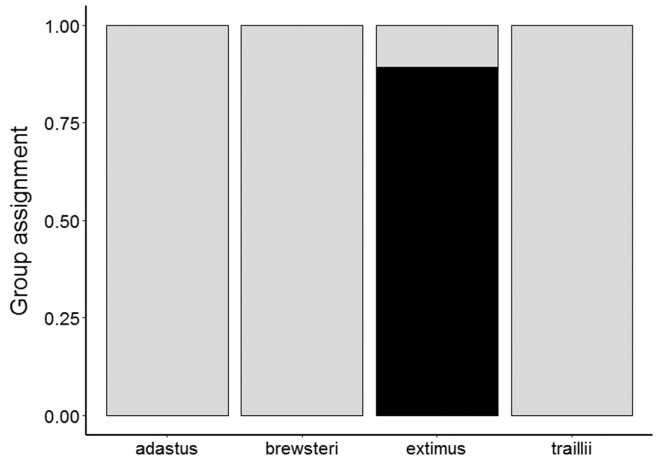

Figure 6. Proportion of song group assignments (black, 'song group 1', gray, 'song group 2') for willow flycatcher subspecies (*Empidonax traillii adastus*, *E. t. brewsteri*, *E. t. extimus*, *E. t. traillii*). Group assignments were based on unsupervised and unbiased classification of song principal components derived from a principal components analysis. Song group 1 was comprised entirely of *E. t. extimus* (89%, n = 33). All *E. t. adastus* (n = 46), *E. t. brewsteri* (n = 32), *E. t. traillii* (n = 26) and 11% of *E. t. extimus* (n = 4) songs grouped into song group 2.

a distinct song category, four (11%) of the *E. t. extimus* songs used in our study grouped with the other subspecies. Of these songs, three were from potential contact zones between *E. t. extimus* and *E. t. adastus* and could have represented birds of the other subspecies or admixed individuals (Kern River, CA (n = 2) and Virgin River at St George, UT (n = 1); Paxton et al. 2008, Theimer et al. 2016, M. J. Whitfield pers. comm.). The remaining individual was recorded well within the *E. t. extimus* range on the Rio Grande River near Elephant Butte Reservoir, NM. Future genomic studies and playback experiments will be important in determining the identity of these individuals.

Song is proposed to be a reproductive isolating mechanism in willow flycatchers (Stein 1963) and was previously found to differ between *E. t. adastus* and *E. t. extimus* (Sedgwick 2001). Our results generally agree with Sedgwick (2001). Although Sedgwick (2001) found differences in song length between *E. t. extimus* and *E. t. adastus* (their Table 2), in our study, song length was not an important variable in our PCA. However our song group that contained only *E. t. extimus* was distinguished by fewer frequency modulations in part 2 of phrase 3, overall lower frequencies of notes in phrases 1 and 2, and lower frequencies at maximum song amplitude,

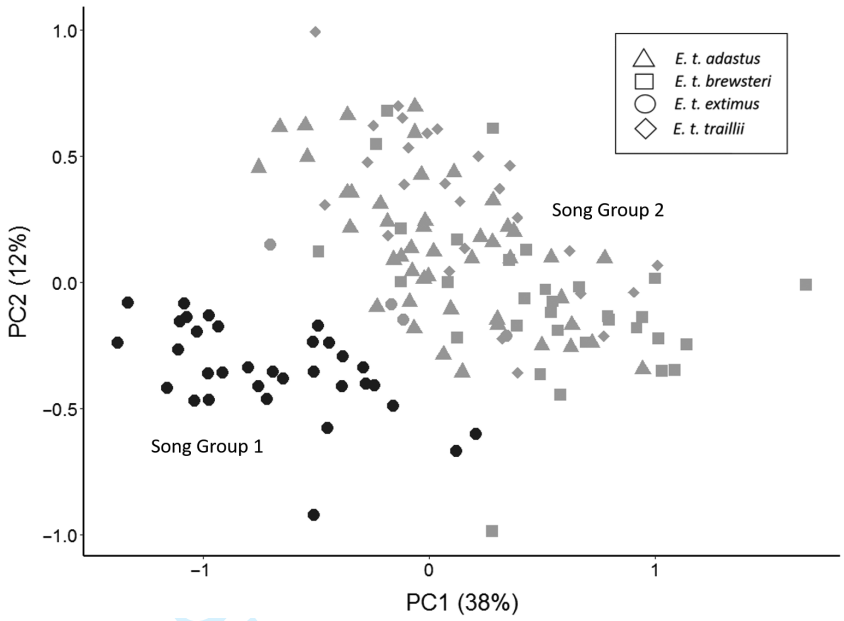

Figure 7. Willow flycatcher song PC1 (38% of variation) and PC2 (12% of variation). Symbols represent putative subspecies identification, based on location of field recordings during the breeding season. Colors represent song group assignment (black, 'song group 1', gray, 'song group 2'), based on unsupervised and unbiased classification of song principal components derived from a principal components analysis. Positive PC1 scores represent higher frequency songs. Positive PC2 scores represent more frequency modulations in Phrase 3.

consistent with the same key differences between *E. t. extimus* and *E. t. adastus* as described by Sedgwick (2001, their Table 2, 3).

Song in tyrannids is innate rather than learned (Kroodsma 1984), so differences in song arise through genetic rather than cultural evolution. Several non-exclusive hypotheses may explain the pattern of song differentiation in willow flycatchers that we documented. Differences among populations could arise through mutations that alter syrinx morphology, neural circuitry or other physical aspects of song (Stein 1963, Isler et al. 1997, Robbins and Stiles 1999, Sedgwick 2001, Podos and Warren 2007). Once established those differences could be maintained through either selection or reduced gene flow that prevents genetic homogenization. *Empidonax* species and subspecies often show strong habitat preferences and generally have geographic ranges that are broadly allopatric (Sedgwick 2000, Johnson and Cicero 2002) and this may reduce gene flow between populations and increase the role of drift (Sedgwick 2001). Of the four willow flycatcher subspecies, *E. t. extimus* arguably inhabits the most distinct habitat, breeding in lower elevation riparian vegetation in the arid southwestern United States. Selection pressure to optimize signal transmission in those hotter and drier habitats may have shaped song structure (Morton 1975, Slabbekoorn 2004, Seddon 2005, Wilkins et al. 2013). In fact, willow flycatchers song appears to be consistent with this hypothesis, as songs from populations in more arid regions (i.e. *E. t. extimus*) are lower in frequency and exhibit fewer frequency modulations (Gonzalez et al. unpubl.). Alternatively, selectively neutral changes in song could be more likely to be established through founder effects and maintained by isolation in southwestern population due to the rarity of appropriate riparian habitat within the broader arid landscape. Whether individuals recognize the vocal differences we documented remains to be tested but such data would be an important step in understanding population divergence in willow flycatchers.

Finally, our results have important implications for taxonomy and conservation of the southwestern populations of the willow flycatcher. In a recent commentary, Zink (2015) called into question the validity of the subspecies designation of the southwestern willow flycatcher, *E. t. extimus*, and therefore its protection under the Endangered Species Act. Several concerns voiced in that paper have been addressed here. Previous studies of plumage either lacked rigorous statistical analyses, used data from live specimens and thus lacked reference specimens for subsequent reanalysis and/or assumed subspecies identity a priori based on presumed subspecies distributions. Our plumage analysis was based on catalogued museum specimens (admittedly with the caveats of using museum specimens listed above) and utilized unbiased classification analyses that did not assume subspecies identification. Likewise, our song analysis was based on publicly available songs from all four subspecies, and our analyses grouped songs independent of any a priori assumptions about subspecies identity. Zink (2015) also proposed adopting Amadon's (1949) '75% rule' to justify considering a subset of a species as a distinct subspecies. This rule requires that '75% of a population effectively must lie outside 99% of the range of other populations for a given defining character or set of characters' (Patten and Unitt 2002, p. 27). Zink (2015) went further and recommended characters 'be nearly

(95%) if not completely diagnosable (Cracraft et al. 1998).' Our unbiased classification analyses of song structure separated songs into two groups, with 89% of putative *E. t. extimus* falling into song group 1 and 100% of three subspecies falling into song group 2, thereby exceeding the '75% rule' and approximating the more rigorous recommendation of Zink (2015). Importantly, this diagnosis was based on song, a trait essential in diagnosing species limits in Tyrannidae (Rheindt et al. 2008, Tobias et al. 2010). Thus our song data support recognition of the southwestern population as a distinct subspecies.

Acknowledgments – We thank the following individuals for providing access to museum collections: Andy Johnson (Univ. of New Mexico), Phil Unitt (San Diego Natural History Museum), Jessika Vasquez (San Bernardino County Museum), Carla Cicero (Museum of Vertebrate Biology, Univ. of California, Berkeley), Peter Konstantindis (Oregon State Univ.), Robert Faucett (Burke Museum), Libby Beckman (Phillip L. Wright Zoologica Museum, Univ. of Montana), Jeff Stephenson (Denver Museum of Nature and Science), Elizabeth Wommack (Univ. of Wyoming Museum of Vertebrates), Eric Rickart (Natural History Museum of Utah). We also thank the Macaulay Library (Cornell Univ.) and xeno-canto.org libraries for generously sharing song recordings. Comments from the Theimer lab group (Northern Arizona Univ.) and C. E. Aslan and S. M. Shuster greatly improved this manuscript. S. I. Gonzalez, R. W. Winton, A. N. B. Smith and D. N. Rakestraw assisted with gathering field recordings. Finally, comments from three anonymous reviewers greatly improved this paper.
Funding – This work was supported by the T&E Inc. Conservation Grant (SMM), NAU John Prather Conservation Award (SMM) and the NAU Landscape Conservation Initiative Fellowship (SMM).
Conflicts of interest – The authors declare no conflict of interest.
Permits – No permits were required for this research.

Author contributions

Sean Mahoney: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Funding acquisition (lead); Investigation (lead); Methodology (equal); Resources (equal); Writing – original draft (lead); Writing – review and editing (equal). **Matthew Reudink**: Conceptualization (supporting); Formal analysis (supporting); Methodology (supporting); Resources (equal); Writing – review and editing (equal). **Bret Pasch**: Conceptualization (equal); Formal analysis (supporting); Methodology (equal); Resources (equal); Writing – original draft (supporting); Writing – review and editing (equal). **Tad Theimer**: Conceptualization (equal); Formal analysis (supporting); Methodology (equal); Resources (equal); Writing – original draft (supporting); Writing – review and editing (equal).

Data availability statement

Spectrophotometry data, specimen photographs, song files used in this study are provided by the authors in the Dryad data repository.

References

Aldrich, J. W. 1951. A review of the races of the Traill's flycatcher. – Wilson Bull. 63: 192–197.
Alonso-Alvarez, C., Doutrelant, C. and Sorci, G. 2004. Ultraviolet reflectance affects male–male interactions in the blue tit *Parus caeruleus ultramarinus*. – Behav. Ecol. 15: 805–809.
Amadon, D. 1949. The seventy-five per cent rule for subspecies. – Condor 51: 250–258.
Armenta, J. K., Dunn, P. O. and Whittingham, L. A. 2008. Quantifying avian sexual dichromatism: a comparison of methods. – J. Exp. Biol. 211: 2423–2430.
Barrowclough, G. F., Cracraft, J., Klicka, J. and Zink, R. M. 2016. How many kinds of birds are there and why does it matter? – PLoS One 11: e0166307.
Bible, J., Datta, S. and Datta, S. 2013. Cluster analysis: finding groups in data. – In: Informatics for materials science and engineering. Butterworth-Heinemann, pp. 53–70.
Bradbury, J. W. and Vehrencamp, S. L. 1998. Principles of animal communication. – Sinauer Assoc.
Brewster, W. 1895. Notes on certain flycatchers of the genus *Empidonax*. – Auk 12: 157–163.
Brock, G., Pihur, V., Datta, S. and Datta, S. 2008. clValid, an R package for cluster validation. – J. Stat. Softw. 25. <http://www.jstatsoft.org/v25/i04>.
Browning, M. R. 1993. Comments on the taxonomy of *Empidonax traillii* (willow flycatcher). – West. Birds 24: 241–257.
Cracraft, J., Feinstein, J. Vaughn, J. and Helm-Bychowski, K. 1998. Sorting out tigers *Panthera tigris*: mitochondrial sequences, nuclear inserts, systematics and conservation genetics. – Anim. Conserv. 1: 139–150.
Cuthill, I. C., Partridge, J. C., Bennett, A. T., Church, S. C., Hart, N. S. and Hunt, S. 2000. Ultraviolet vision in birds. – Adv. Study Behav. 29: 159–214.
Doucet, S. M. and Hill, G. E. 2009. Do museum specimens accurately represent wild birds? A case study of carotenoid, melanin and structural colours in long-tailed manakins *Chiroxiphia linearis*. – J. Avian Biol. 40: 146–156.
Dunn, C. 1974. Well separated clusters and fuzzy partitions. – J. Cybernet. 4: 95–104.
Eaton, M. D. 2007. Avian visual perspective on plumage coloration confirms rarity of sexually monochromatic North American passerines. – Auk 124: 155–161.
Griffith, S. C., Ornborg, J., Russell, A. F., Andersson, S. and Sheldon, B. C. 2003. Correlations between ultraviolet coloration, overwinter survival and offspring sex ratio in the blue tit. – J. Evol. Biol. 16: 1045–1054.
Handl, J., Knowles, J. and Kell, D. B. 2005. Computational cluster validation in postgenomic data analysis. – Bioinformatics 21: 3201–12.
Haig, S. M., Beever, E. A., Chambers, S. M., Draheim, H. M., Dugger, B. D., Dunham, S., Elliot-Smith, E., Fontaine, J. B., Kesler, D. C., Knaus, B. J. and Lopes, I. F. 2006. Taxonomic considerations in listing subspecies under the US Endangered Species Act. – Conserv. Biol. 20: 1584–1594.
Hill, G. E. 2006. Environmental regulation of ornamental coloration. – In: Hill, G. E. and McGraw, K. J. (eds), Bird coloration, Vol. 1. Mechanisms and measurements. Harvard Univ. Press, pp. 507–560.
Hunt, S., Cuthill, I. C., Bennett, A. T. D. and Griffiths, R. 1999. Preferences for ultraviolet partners in the blue tit. – Anim. Behav. 58: 809–815.

Jeon, J. Y. and Hong, J. Y. 2015. Classification of urban park soundscapes through perceptions of the acoustical environments. – Land. Urban Plan. 141: 100–111.

Johnson, N. K. and Cicero, C. 2002. The role of ecologic diversification in sibling speciation of *Empidonax* flycatchers (Tyrannidae): multigene evidence from mtDNA. – Mol. Ecol. 11: 2065–2081.

Kassambara, A. and Mundt, F. 2017. Package 'factoextra'. Extract and visualize the results of multivariate data analyses. 76 p.

Kroodsma, D. E. 1984. Songs of the alder flycatcher *Empidonax alnorum* and willow flycatcher *Empidonax traillii* are innate. – Auk 101: 13–24.

Isler, M. L., Isler, P. R. and Whitney, B. M. 1997. Biogeography and systematics of the *Thamnophilus punctatus* (*Thamnophilidae*) complex. – In: Remsen Jr., J. (ed.), Studies in neotropical ornithology honoring ted parker. American Ornithologists' Union, pp. 355–382.

Limbourg, T., Mateman, A. C., Andersson, S. and Lessells, C. M. 2004. Female blue tits adjust parental effort to manipulated male UV attractiveness. – Proc. R. Soc. B 271: 1903–1908.

Lovette, I. J. and Bermingham, E. 1999. Explosive speciation in the New World *Dendroica* warblers. – Proc. R. Soc. B 226: 1629–1636.

Lucek, K., Kristjánsson, B. K., Skúlason, S. and Seehausen, O. 2016. Ecosystem size matters: the dimensionality of intralacustrine diversification in Icelandic stickleback is predicted by lake size. – Ecol. Evol. 6: 5256–5272.

Maia, R., Eliason, C. M., Bitton, P.-P., Doucet, S. M. and Shawkey, M. D. 2013. Pavo: an R package for the analysis, visualization and organization of spectral data. – Meth. Ecol. Evol. 4: 906–913.

Mallet, J. 2005. Hybridization as an invasion of the genome. – Trends Ecol. Evol. 20: 229–237.

Martens, J., Eck, S., Päckert, M. and Sun, Y.-H. 2003. Methods of systematic and taxonomic research on passerine birds: the timely example of the *Seicercus burkii* complex (Sylviidae). – Bonn. Zool. Beitr. 51:109–118.

Mason, N. A. and Unitt, P. 2018. Rapid phenotypic change in a native bird population following conversion of the Colorado Desert to agriculture. – J. Avian Biol. 49: jav-01507.

Mayr, E. 1942. Systematics and the origin of species. – Columbia Univ. Press.

McNett, G. D. and Marchetti, K. 2005. Ultraviolet degradation in carotenoid patches: live versus museum specimens of wood warblers (Parulidae). – Auk 122: 793–802.

Montgomerie, R. 2006. Analyzing colors. – In: Hill, G. E. and McGraw, K. J. (eds), Bird coloration, Vol. 1. Mechanisms and measurements. Harvard Uni. Press, pp. 90–147.

Morton, E. S. 1975. Ecological sources of selection on avian sounds. – Am. Nat. 109: 17–34.

Oberholser, H. C. 1918. New light on the status of *Empidonax traillii* (Audubon). – Ohio J. Sci. 18: 85–98.

Oberholser, H. C. 1932. Descriptions of new birds from Oregon, chiefly from the Warner Valley region. Sci. Publ. Cleveland Museum Nat. Hist. 4: 1–12.

Oberholser, H. C. 1947. A new flycatcher from the western United States. – Proc. Biol. Soc. Washington 60: 77–78.

Ödeen, A. and Håstad, O. 2003. Complex distribution of avian color vision systems revealed by sequencing the SWS1 opsin from total DNA. – Mol. Biol. Evol. 20: 855–861.

Päckert, M., Martens, J., Kosuch, J., Nazarenko, A. A. and Veith, M. 2003. Phylogenetic signal in the song of crests and kinglets (Aves: Regulus). – Evolution 57: 616–629.

Patten, M. A. and Unitt, P. 2002. Diagnosability versus mean differences of sage sparrow subspecies. – Auk 119: 26–35.

Paxton, E. H. 2000. Molecular genetic structuring and demographic history of the willow flycatcher *Empidonax traillii*. – MS thesis, Northern Arizona Univ., Flagstaff, AZ.

Paxton, E. H., Sogge, M. K., Theimer, T. C., Girard, J. and Keim, P. 2008. Using molecular markers to resolve a subspecies boundary: the northern boundary of the southwestern willow flycatcher in the Four Corner states. – U.S. Geological Survey Open File Report 2008-1117.

Paxton, E. H., Sogge, M. K., Koronkiewicz, T. J., McCleod, M. A. and Theimer, T. C. 2010. Geographic variation in the plumage coloration of willow flycatchers *Empidonax traillii*. – J. Avian Biol. 41: 128–138.

Phillips, A. R. 1948. Variation in *Empidonax traillii*. – Auk 65: 507–514.

Pihur, V., Datta, S. and Datta, S. 2009. *RankAggreg*, an R package for weighted rank aggregation. – BMC Bioinform. 10: 62.

Podos, J. and Warren, P. S. 2007. The evolution of geographic variation in birdsong. – Adv. Study Behav. 37: 403–458.

Pohland, G. and Mullen, P. 2006. Preservation agents influence UV-coloration of plumage in museum bird skins. – J. Ornithol. 147: 464–467.

Prescott, D. R. 1987. Territorial responses to song playback in allopatric and sympatric populations of alder *Empidonax alnorum* and willow *E. traillii* flycatchers. – Wilson Bull. 99: 611–619.

Prum, R. O. 2010. The Lande–Kirkpatrick mechanism is the null model of evolution by intersexual selection: implications for meaning, honesty and design in intersexual signals. – Evolution 64: 3085–3100.

Reudink, M. W., Marra, P. P., Boag, P. T. and Ratcliffe, L. M. 2009. Plumage coloration predicts paternity and polygyny in the American redstart. – Anim. Behav. 77: 495–501.

Rheindt, F. E., Norman, J. A., Christidis, L. 2008. DNA evidence show vocalizations to be a better indicator of taxonomic limits than plumage patterns in *Zimmerius* tyrant-flycatchers. – Mol. Phylogenet. Evol. 48: 150–156.

Robbins, M. B. and Stiles, F. G. 1999. A new species of pygmy-owl (Strigidae: *Glaucidium*) from the Pacific slope of the northern Andes. – Auk 116: 305–315.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. – J. Comp. Appl. Math. 20: 53–65.

Seddon, N. 2005. Ecological adaptation and species recognition drives vocal evolution in neotropical suboscine birds. – Evolution 59: 200–215.

Sedgwick, J. A. 2000. Willow flycatcher *Empidonax traillii*. – In: Poole, A. and Gill, F. B. (ed.), The Birds of North America, Inc., Philadelphia, PA, USA.

Sedgwick, J. A. 2001. Geographic variation in the song of willow flycatchers: differentiation between *Empidonax traillii adastus* and *E. t. extimus*. – Auk 118: 366–379.

Shutler, D. and Weatherhead, P. J. 1990. Targets of sexual selection: song and plumage of wood warblers. – Evolution 44: 1967–1977.

Slabbekoorn, H. 2004. Singing in the wild: the ecology of birdsong. – In: Marler, P. and Slabbekoorn, H. (eds), Nature's music: the science of birdsong. Elsevier Academic Press, pp. 178–205.

Stein, R. C. 1958. The behavioral, ecological and morphological characteristics of two populations of the alder flycatcher: *Empidonax Trailli* (Audobon) (no. 371). – Univ. of the State of New York, State Education Dept, NY, USA.

Stein, R. C. 1963. Isolating mechanisms between populations of Traill's flycatchers. – Proc. Am. Phil. Soc. 107: 21–31.

Theimer, T. C., Smith, A. D., Mahoney, S. M. and Ironside, K. E. 2016. Available data support protection of the Southwestern willow flycatcher under the Endangered Species Act. – Condor Ornithol. Appl. 118: 289–299.

Tobias, J. A., Seddon, N., Spottiswoode, C. N., Pilgrim, J. D., Fishpool, L. D. and Collar, N. J. 2010. Quantitative criteria for species delimitation. – Ibis 152: 724–746.

Toews, D. P., Taylor, S. A., Vallender, R., Brelsford, A., Butcher, B. G., Messer, P. W. and Lovette, I. J. 2016. Plumage genes and little else distinguish the genomes of hybridizing warblers. – Curr. Biol. 26: 2313–2318.

Unitt, P. 1987. *Empidonax traillii extimus*: an endangered subspecies. – West. Birds 18: 137–162.

USFWS (U.S. Fish and Wildlife Service) 1995. Final rule determining endangered status for the southwestern willow flycatcher. – Federal Register 60: 10694–10715.

Vehrencamp, S. L., Ritter, A. F., Keever, M. and Bradbury, J. W. 2003. Responses to playback of local vs. distant contact calls in the orange-fronted conure, *Aratinga canicularis*. – Ethology 109: 37–54.

Vorobeyev, M., Osorio, D., Bennett, A. T. D., Marshall, N. J. and Cuthill, I. C. 1998. Tetrachromacy, oil droplets and bird plumage colours. – J. Comp. Phys. A 183: 621–633.

Wilkins, M. R., Seddon, N. and Safran, R. J. 2013. Evolutionary divergence in acoustic signals: causes and consequences. – Trends Ecol. Evol. 28: 156–166.

Zink, R. M. 2015. Genetics, morphology and ecological niche modeling do not support the subspecies status of the endangered southwestern willow flycatcher *Empidonax traillii extimus*. – Condor Ornithol. Appl. 117: 76–86.

Zink, R. M. and Johnson, N. K. 1984. Evolutionary genetics of flycatchers. I. Sibling species in the genera *Empidonax* and *Contopus*. – Syst. Biol. 33: 205–216.

Author Queries

JOB NUMBER: 2621

JOURNAL: OIK_JAB

Q1 Please provide the page range or article ID for reference 'Brock et al. 2008'.

Q2 Please update the reference 'Kassambara and Mundt 2017'.