Revision round #1
26 Aug 2024

Decision for round #1 : Revision needed

Dear Michael D Greenfield,

We thank you and the reviewers for your thoughtful comments on our manuscript. We have carefully considered all the feedback and believe the suggestions have significantly improved the quality of our work. Below, we provide point-by-point responses to the reviewers' comments and highlight the corresponding changes made to the manuscript.

In addition to the revisions made in response to the reviewers' comments, we have implemented two additional modifications based on the advice of Pierre de Villemereuil, a specialist in biostatistics for ecology, who reviewed this manuscript as part of my PhD evaluation. First, the method for estimating phylogenetic heritability was adjusted to account for the non-linearity of the model. The previous equation, often used to estimate phylogenetic heritability in GLMMs, is valid but relies on a linear approximation. Second, the statistical model of the Phylogenetic Path Analysis was refined by applying a log() transformation to the response variable, ensuring a better model fit and enhancing the robustness of the estimates. These minor adjustments in the analyses did not affect the results and therefore had no impact on the main messages of the manuscript.

We sincerely appreciate your consideration of our revised manuscript, and we hope it meets your expectations. Please find the updated version of our manuscript in the OSF preprints database: “https://doi.org/10.31219/osf.io/z5bgy_v2”, along with our detailed responses to the reviewers' comments and questions, provided below in blue text. Data, scripts, supplementary material and a version of the manuscript with highlighted corrections can be found at the following private OSF link:
“https://osf.io/k7jr6/?view_only=a0fa474ccea24ba19e168e2bb08969d1”

Corresponding author, Erwan Harscouet-Commecy

erwan.harscouet@protonmail.com

Review by anonymous reviewer 1, 22Aug 2024 20:02

Summary:

The authors have completed a study that first defines two measures of vocalizations in a group of birds, and employ phylogenetic comparative analyses to determine which potential species variables might affect vocalizations across species. The two variables the authors generated to describe vocal diversity are song diversity and song composition. The authors then determine if reproductive system, social system/size, and phylogenetic history influence the differences across species. The authors find that social system and size affect both of the variables, however, the mating system does not significantly affect either of the two measures of vocalizations. The authors interpret these results as social system being of considerable influence on the evolution of vocalizations in the group of species they have studied.

General Comments:

The authors have generated an especially interesting dataset to answer an important question in the evolution of communication and signal diversity. To test how communication changes across species, the authors employ appropriate comparative analyses and are thorough in their analysis. Most of the changes I have suggested are (hopefully) simple but I have two larger concerns I was hoping the authors might address. First, the authors break species down into categories of colony size rather than use colony size as a numerical variable in the analysis. I was hoping the authors could provide an argument for the utility of examining colony size as a category rather than a numerical variable. While some weaver species have exceedingly large colonies, a log transformation of colony size might prevent overly strong influence of certain species in a comparative analysis.

Additionally, I would suggest the authors increase the depth of explanation of the song composition explanation in the methods. The authors have created a set of vocalization groups/types that may be present in multiple species, and having increased information about how these were defined, what the definitions are, and perhaps a visual example of 2-3 would be considerably helpful.

We thank you for your reading, comments and feedback. Concerning your broader concerns, we address them below, while the other comments are answered individually later.

First, we explain here in more detail why we coded the colony size as a categorical variable. We used the literature to estimate this variable for each species, and because for around 25% of the species included in the study, colony size was estimated by 3 records or less (see Oschadleus, 2020), calculating a mean colony size of such small sample would give a poor estimation of the true mean. Note that for most of those species, the records have been corroborated by qualitative information on the colony size (see del Hoyo et al., 2010). We therefore chose to qualitatively quantify colony size because treating colony size as a continuous variable might give the false impression that this variable was precisely measured. However, your comment is relevant, and to assess the robustness of our results, we performed additional analyses using the available quantitative data, log-transformed as suggested. Our results do not change: see Fig. 1 below.

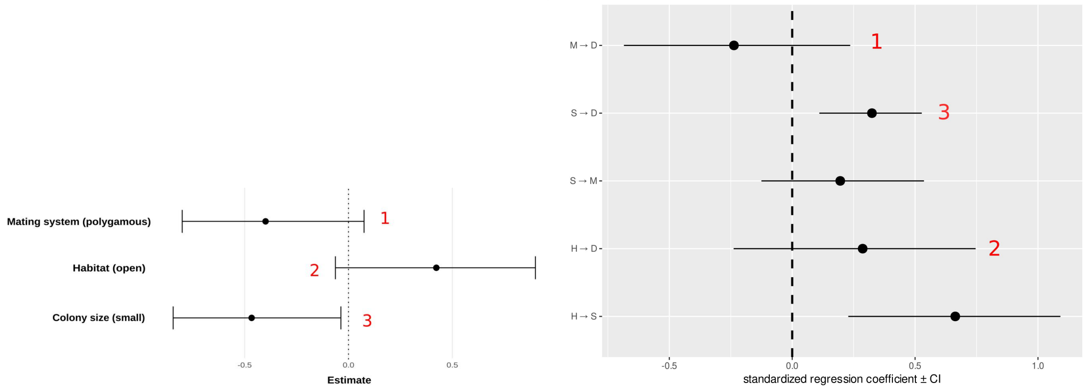

Figure 1: Comparison between the Phylogenetic Path Analysis with colony size as categorical variable (left) and the Phylogenetic Path Analysis with colony size as a continuous variable (right)

M-->D corresponds to the effect of mating system on song diversity (1).
S-->D corresponds to the effect of sociality on song diversity (3).
H-->D corresponds to the effect of habitat on song diversity (2).
S-->M and H-->S corresponds to the effect of sociality on mating system and habitat on sociality.

Comparing these two analyses, the effects are qualitatively similar:
-There is a slightly stronger effect of sociality.
-The 95% confidence interval is narrower for sociality.
-The effects of habitat and mating system are slightly weaker.

For these reasons, we think that treating colony size as a categorical is relevant. We mention in the ms that the results are comparable when colony size is treated differently (l.382-385) and could provide the result in the ESM if asked.

Second, your suggestion of providing visual examples of categories is indeed highly relevant and this was the purpose of the supplementary information submitted with the main text of the manuscript. We are afraid that you did not access to it, for unknown reasons. One of the supplementary file includes a table in which the 59 categories of syllables are defined and illustrated. We hope you will find this table useful, and we remain open to any further comments about it.

Specific Comments

Abstract: None

Introduction:

- Lines 62-63: “unique elements that constitute a signal enables in addition to study signal composition.” The wording at the end of the sentence seems to have omitted a phrase or a clause. Please reword.

We removed the sentence to avoid introducing the concept of signal composition before properly defining it. We now define it later in the text (l.87)

- Lines 66-68: The authors argue that “similarities in signals across different evolutionary contexts have been insufficiently studied”. However, one example of similarity in signals is mimicry, and mimicry is a foundational field of study in ecology and evolution. I would suggest the authors revise the sentence.

The sentence was reworded: “Similarities in signals have been mainly studied in the context of mimicry (Raguso, 2008), but they also occur in other evolutionary contexts that require further investigation to understand why some species share certain elements while others do not.” (l.66)

- Lines 87-89: The authors are describing signal composition, and to increase the clarity of the sentence they might add an example of describing signal composition. For instance, peak frequency or frequency modulation or a parameter that is especially salient for the current study.

Yes, we thank you for this suggestion. We detailed the definition and added an exemple (l.88): “Song composition, on the other hand, relies on the identification of discrete categories based on the sound characteristics of the acoustic units that make up a song. These acoustic units (or syllables) are grouped into categories according to their acoustic properties (Kershenbaum et al., 2014). For instance, in their study, Price and Lanyon (2002) define different categories based on specific acoustic criteria, such as trills, clicks, rattles, or whistles. The authors then explore how these sound categories are distributed across the phylogeny. Often, it is assumed that the observed acoustic properties stem from different sound production mechanisms used by the signallers. This approach enables comparisons across species by identifying shared or distinct acoustic features within a given taxonomic group (Odom et al., 2021; Sung et al., 2005).”

Materials & Methods:

- Lines 228-230: The syllable categories is an interesting measure but one question would be the sound correlation among syllables produced by different species.

We are not entirely sure we understand what you mean by 'sound correlation'. Are you asking whether different syllables within a category sound similar? In our study, we focus on general acoustic features, meaning that each category encompasses a range of syllables sharing the same predefined acoustic characteristics, despite some variation in other parameters. Therefore, even within the same species, two different syllables can belong to the same category if they follow similar acoustic patterns. For example, the category 'rapid succession of pulses' may include pulses at 3kHz, 4kHz, or 8kHz, all characterized by their rapid emission in sequence. Importantly, the definition of the 59 syllable categories aimed to capture the full repertoire while keeping within-category variation relatively narrow. We hope you now have access to the table detailing how the different categories were constructed.

- Lines 238-239: For species where multiple individuals are measured, is song diversity averaged across individuals (after taking an average per individual if there are multiple songs recorded per individual)?

No, we did not average song diversity across individuals, but we agree it was maybe not clear enough in the previous version of the ms. It is now mentioned more clearly in the M&M (l.246) “Song diversity was estimated in all species (95 species) retaining individual values without averaging them at the species level” and (l.304) “Because we have several individuals for each species we included the species as random effect”. And at the individual level, when several songs were analysed per individual, we specified (l.245) “song diversity for each individual was calculated as the mean for all the produced songs by a given individual.”

- Lines 261-262: The authors state, “Accumulation curves could not be generated for song diversity, as syllables are redefined for each song”. While true, the authors might perform a simple PGLS to determine if the number of individuals recorded per species predicts overall species song diversity. Such an analysis would help confirm that song diversity is not tracking sampling intensity in each species. Alternatively, the authors might down sample to the species with the smallest sampling intensity and determine if there is a relationship between their predictor variables and the dependent variable of song diversity.
Thank you for this insightful comment. We understand the interest in exploring how sample size might influence the estimation of species means. However, we did not perform such an analysis because, theoretically, adding more individuals should primarily increase the precision of the estimated mean, that is, it should reduce the difference between the estimated mean and the true mean, without introducing a systematic bias in one direction or the other. Any variation observed with an increasing sample size would be expected to occur by chance rather than reflecting a consistent trend. The same reasoning applies to the down-sampling suggestion: adjusting the sample size would mainly affect the dispersion of the estimated mean around the true mean, rather than shifting the mean itself. That said, we truly appreciate your suggestion and remain open to further discussion if you believe there are additional nuances we might have overlooked.

- Lines 265-267: Employing mating system as a proxy for sexual selection is a good idea, but one concern is the variation in polygamy. Is there considerable variation in the number of mates a focal individual might have across species? For instance, do some species maintain 2-3 mates per individual, whereas in others there might be 7-10 mates per individual?
That’s a good point. In weaverbirds, some polygamous species have males mating with up to eight females, while in others, the maximum is four. As you pointed out, this suggests that the variance in male mating success may differ between species. However, the variation in mating success variance among most polygamous species is unknow (no field work has been performed to quantify it) and we expect polygamous species to cluster distinctly from monogamous species in terms of the intensity of sexual selection.

Results:

- Lines 354-356: The authors find that colony size affects increased song diversity. The colony size variable in this analysis has been compressed from the four levels into two categories (as I understand it). If the authors were to instead repeat the analysis where you have ‘colonial’ vs ‘noncolonial’ categories do the results remain qualitatively identical? In other words, if the ‘small’ colony size was added to the rest of the colonial species.
Yes, thank you for pointing out this aspect of the analysis. To test the robustness of this analysis and verify whether there is a difference in song diversity between ‘colonial’ and ‘non-colonial’ species, we reran the Phylogenetic Path Analysis using the suggested categories. The results are presented in Figure 2 below. The results of both analyses are very similar, with the effect of colony size being slightly stronger in the ‘colonial’ vs. ‘non-colonial’ analysis.

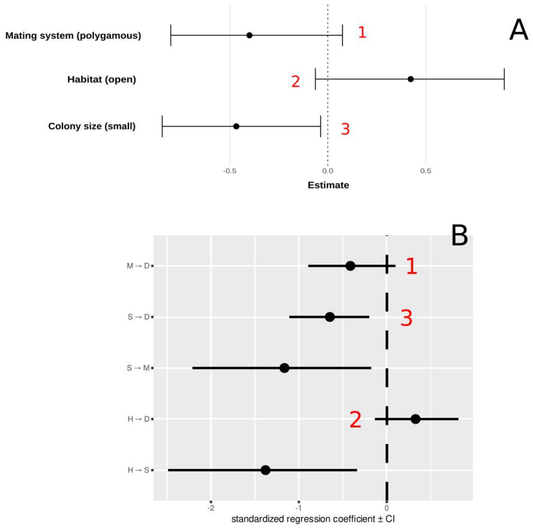

Figure 2: Comparison between (A) the PPA with 2 colony size categories (‘small’ and ‘large’) and (B) the PPA with 2 categories of species ‘colonial’ vs ‘noncolonial’. S-->D correspond to the effect of sociality on song diversity (3). M-->D correspond to the effect of mating system on song diversity (1) and H-->D the effect of habitat on song diversity (2). S-->M and H-->S corresponds to the effect of sociality on mating system and habitat on sociality.

- Line 363: Remove the phrase “results showed that”.

Done

Discussion:

- Line 423: The authors state that song in weaverbirds has hardly been studied, however, several studies have documented the acoustic properties and potential function of vocalizations in these species. Thus I would suggest the authors reword the sentence to say something along the lines of “The structure and diversity of functions has not been investigated across species”.

We agree and followed the suggestion. The phrase is now reworded in: “Weaverbirds are widely known for their bustling colonies, where they can be observed parading around their intricately woven nests. While the acoustic properties and potential functions of their vocalizations have been studied in some species (Collias, 2000; Craig, 1976), the structure and diversity of their songs have not been investigated across species.” (l.442)

- Lines 484-488: The authors presented a plausible explanation of their results regarding sexual selection and song parameters in this section (and the paragraph overall). The authors mention that sexual selection might act on other aspects of song, such as the ability to repeat a note. Sexual selection might also push species towards producing notes/vocalizations at the performance limit given motor constraints. One study to consider citing here is https://royalsocietypublishing.org/doi/10.1098/rsbl.2008.0626.

Thank you for your comment. We agree that mentioning performance limits in song production within the context of sexual selection is relevant here. We have therefore modified the sentence as follows: “For instance, the repetition of the same syllable (Sierro et al., 2023) or the production of syllables or notes at a performance limit (DuBois et al., 2009) can indicate an individual's neuromotor capabilities or quality, which might be associated with its fighting capacity.” (l.509)

- Line 525: Slightly modify “play a role in the variance” to “play a role in explaining the variance”.

Done

Citations:

In line 113 the authors refer to a “Santos et al, 2023” but the methods list has a “dos Santos et al, 2023” study. Are these the same? Please check.

It was a mistake, and we now cite correctly this study.

Figures & Tables:

The figures and tables in the main manuscript are well constructed and informative.

We thank you for this comment!

Supplementary material:

Was not able to access the supplementary material.

We are sorry about this. The supplementary material is located in the folder “OSF_weaver_song_PCJ_R1” from OSF, in a subfolder named “supplementary_information”. As provided in the submission, you can access this folder using the following private link:
https://osf.io/k7jr6/?view_only=a0fa474ccea24ba19e168e2bb08969d1
The file can be downloaded from the 'Files' tab located in the header menu.

Review by anonymous reviewer 2, 22 Aug 2024 16:06

Title and abstract
Does the title clearly reflect the content of the article?
[XX] Yes, [ ] No (please explain), [ ] I don't know

Does the abstract present the main findings of the study?
[XX] Yes, [ ] No (please explain), [ ] I don’t know

Introduction
Are the research questions/hypotheses/predictions clearly presented?
[XX] Yes, [ ] No (please explain), [ ] I don’t know

Does the introduction build on relevant research in the field?
[XX] Yes, [ ] No (please explain), [ ] I don’t know

Materials and methods
Are the methods and analyses sufficiently detailed to allow replication by other researchers?
[ ] Yes, [XX] No (please explain), [ ] I don’t know --> see comments to authors regarding 'composition' and inter-observer reliability

Are the methods and statistical analyses appropriate and well described?
[XX] Yes, [ ] No (please explain), [ ] I don’t know

Results
In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)?
[ ] Yes, [ ] No (please explain), [XX] I don’t know

Are the results described and interpreted correctly?
[XX] Yes, [ ] No (please explain), [ ] I don’t know

Discussion
Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument?
[ ] Yes, [XX] No (please explain), [ ] I don’t know --> see comments to authors regarding alternative explanation of song complexity potentially driving social complexity

Are the conclusions adequately supported by the results (without overstating the implications of the findings)?
[ZZ] Yes, [ ] No (please explain), [ ] I don’t know

Review of Manuscript:

This manuscript describes the complexity of song in weaverbirds and tests different factors that might explain that complexity. A strength of the study is that the authors tested multiple hypotheses – too many studies only test the Social Complexity Hypothesis to test whether larger groups or groups with more complicated social structures have larger signal repertoires. Another strength is that the authors looked at two metrics of song complexity – diversity and composition. The authors found that phylogenetic relationship and social complexity were strongly associated with song complexity, whereas sexual selection was not.

Although there is much to this manuscript that I like, I do have several concerns and comments. Two I see as really important to address. First, I was never completely clear on what ‘composition’ of the songs was – I have a general idea but am not sure. For sure, the authors need to define the term explicitly and early in the manuscript. The Kershenbaum et al. 2016 paper “Acoustic sequences in non-human animals: a review and prospectus” (Biol Rev 91 doi: 10.1111/brv.12160) seems relevant to this term. Second, the authors really need to report inter-observer reliability statistics for the coding of song elements (see lines 230-232). Also, it would be good to have inter-observer reliability statistics on the colony size differences variables they used (see lines 284-286) – I could see this potentially being a bit subjective.

We thank you for your reading and comments and feedback. Concerning your larger concerns, we answer to them hereafter and for the other comments we address them individually later.

****

Song composition definition

****

Concerning the definition of song composition, we agree that a more precise definition is needed. Thank you for suggesting to add a mention to Kershenbaum et al. 2016 paper. We have now clarified the definition of song composition in the introduction, introduced it earlier, and made additional changes to improve the fluency of the text.

a/ We removed the sentence: “Investigating these unique elements, that constitute a signal, enables us to study signal composition” to avoid introducing the concept of signal composition before properly defining it. We now define it later in the text.

b/ In paragraph 3, we define both signal diversity and signal composition in a more general way, as commonly found in the literature. It is in this section that we now provide a detailed definition of composition and include an example (l.88): “Song composition, on the other hand, relies on the identification of discrete categories based on the sound characteristics of the acoustic units that make up a song. These acoustic units (or syllables) are grouped into categories according to their acoustic properties (Kershenbaum et al., 2014). For instance, in their study, Price and Lanyon (2002) define different categories based on specific acoustic criteria, such as trills, clicks, rattles, or whistles. The authors then explore how these sound categories are distributed across the phylogeny. Often, it is assumed that the observed acoustic properties stem from different sound production mechanisms used by the signallers. This approach enables comparisons across species by identifying shared or distinct acoustic features within a given taxonomic group (Odom et al., 2021; Sung et al., 2005).” At this point, it seems important to define these terms in relation to the general literature rather than specifically within the context of our study, and the various articles cited in the following paragraphs do not necessarily measure diversity and composition in exactly the same way as we do.

c/ In the last paragraph of the introduction, we explain how song diversity and song composition are estimated in our study. We have not changed this paragraph.

Additionally, to facilitate the flow of reading, we made minor adjustments to the text. In the previous version of the manuscript, in paragraph 1, to illustrate signal diversity, we referred to acoustic signal diversity (lines 56–59). Then, in paragraph 2, we discussed signals in a more general sense (not only acoustic but also visual and chemical signals), and it was only in paragraph 3 that we focused specifically on acoustic signals. To avoid confusion caused by the example in paragraph 1, we have now replaced the type of signal used to illustrate signal diversity (l.56–60) as follows: “For example, some species of birds, such as the spot-breasted oriole (*Icterus pectoralis*), produce gestural displays with low signal diversity, while others, like the shiny cowbird (*Molothrus bonariensis*), exhibit a high diversity, with a wide range of variations (Miles et al., 2017).”

****

Inter-observer reliability: annotation

****

We agree with the reviewer that inter-observer reliability in songs and spectrograms annotation is a really important point, that we fully considered during the study and explain here how we dealt with this.

Assigning syllables to categories was first performed by checking visually whether the shape and key parameters of the sound was corresponding to the general shape of the category previously defined.

Second, although it was not specified in the method, we paid careful attention during annotation processes to maximise inter-observer reliability:

a/ Each of the four annotators who contributed to the sound analysis and spectrogram annotation re-annotated (blindly) a subset of spectrograms previously annotated by others (between 30 and 40 records) to ensure that everyone applied the same annotation criteria, aiming for high consistency across annotators. Annotations were only validated when there was strong agreement among all annotators.

b/ Each of the four annotators repeated their own annotations on a sample (around 10%) of their previous work to ensure the repeatability of the measurements. This exercise aimed to verify that each annotator was consistent in their own measurements.

c/ For each annotator, a portion of their annotations (between 5% and 10%) was double-checked by a fifth person (Fanny Rybak) specialized in birdsong, once again to ensure consistency across annotators.

d/ To limit residual observer biases, each person annotated a range of species, so that most species were annotated by three or four people.

Following these principles, we are very confident in the efficiency and minimisation of subjectivity. This important information is now given in ESM.

****
Inter-observer reliability: colony size
****

Regarding inter-observer reliability statistics on colony size differences, your comments show that there was lack of clarity in our text regarding the coding of colony size differences for performing the statistical model which tested the relationships between song composition and predictor variables.

In this analysis, we aimed to evaluate how differences in ecology and life history traits can predict differences in song composition. Thus for each pair of species, the differences between the characteristics of their song composition and those of their mating system, habitat, and colony size were calculated (see l.283-298). The 'distance in colony size' between a pair of species was obtained considering the number of steps separating the levels of colony size (The 4 levels were “solitary “: 1 to 2 individuals, “small” : 2 to 5 individuals, “medium”: 5 to 15 individuals, “large” : more than 15 individuals, see l.277-278) of one species and the other. There is no room for subjectivity as when 2 species are both at the same level, their difference is coded as ‘comparable’, when they separated by one step, their difference is coded as ‘different’, when they were separated by 2 steps, their difference was coded as ‘very different’ and when they were separated by 3 steps, their difference was coded as ‘extremely different’.

We added a sentence in the manuscript to clarify the construction of the variable : “Thus, the difference in colony size between a pair of species increases with the number of steps separating the colony size levels of one species from the other.” (l.294-295)

I raise the additional comments here by line number and not in any order of importance.

- Line 120-141: Is there any reason to believe that territory or home range densities might explain some of the variation in songs? This is one of the metrics of social complexity that is rarely studied – Terry Ord has published a couple of studies on visual displays in lizards and how they relate to spatial variation and territory densities.
This is an interesting point. Indeed, in the paper “Ecology and signal evolution in lizards” by Terry Ord et al., the authors found that the higher the density (smaller home range), the more complex the signal displays are. They explained that this is probably because, in these species, there is more male-male competition, and it is important to accurately assess the quality of competing males to avoid the costs associated with fights.

This would be an interesting question to explore, but unfortunately, we don't have such data.

- Line 177-181: Both of these definitions should come earlier in the paper, I think, and the one for ‘composition’ may need to be spelled out more thoroughly.
We agree and we modified the introduction accordingly, as explained in detail above.

- Line 201 and 205: I suggest “The song data were collected…” and “Sennheiser microphone (MKH70 with K6 power module)…”
We modified the sentences: “The song data were collected from online sound archives […]” (l.208) and “Recodings in the field were made with a Sennheiser microphone […]” (l.211).

- Line 214-216: The average number of songs recorded per individual seems quite low and I wondered if this might result in problematic estimates of repertoire size/song diversity.
You are right to point out the sample size. In our study, 79% of the species have 3 individuals or more, and 40% have 5 individuals or more. This might seem low, when comparing to studies of bird song at a specific scale. However, compared to other studies on similar topic at a multi-species scale, our dataset is very respectable. See some examples below:

Example 1:
"We collected 1 to 5 (median = 4, mean +/- SD = 3.6 +/- 1.5) recordings of adult male song for each species" about song frequency (Mikula et al., 2021 in Ecology Letters). They do not specify how many individuals were sampled.

Example 2:
"Of the 5,100 passerine bird species for which we found phylogenetic data, we were able to download and extract song data for 578 (mean number of songs per species: 10.26, standard deviation: 16)" about peak frequency and standard deviation of frequency (Pearse et al., 2018 in Evolution). The authors do not specify a minimum number of songs per species, but the value of variance strongly suggests that for several species there was only few songs per species.

Examples 1 and 2 have similar sample sizes per species than in our study, even lower, as they only refer to songs per species without specifying the number of individuals. In our study, we have 3,577 songs for 95 species, corresponding to a mean of 38 songs per species, 4 times more than in Pearse et al. (2018). However, both studies include more species than ours (6 times more for Pearse et al. (2018) and 50 times more for Mikula et al. (2021)).

Example 3:
"When possible, we collected 10 song recordings from each species" (Mejias et al., 2020 in Evolution) about 4 song length and frequency parameters for 50 species.

Example 4:
"We collected a total of 267 audio recordings in .wav format (194 males and 73 females), corresponding to 64 and 38 taxa for males and females, respectively" (Beco et al., 2021 in Evolution) about a dozen parameters.

We acknowledge that our dataset is not huge, but it is comparable to other good studies published in reputable journals.

- Line 219-220: On what data is the 1.5 sec criterion based? If two vocal elements are 1.4 sec apart from one another, that seems like a long time to me to consider them to be in the same song, but it is also quite possible I simply don't know the songs of these species well enough or don't understand the criterion used here correctly.
We set the 1.5 sec criterion based on our previous works on song analysis in other species (common skylark and european robin for example), and on preliminary analyses performed on a subset of recordings of several species of weavers. In those samples, the distribution of silence intervals between sound units was bimodal with a local minimum value between the 2 modes of around 1.5 sec.

- Line 237-239: How were specific individuals known in these multi-individual recordings given – I assume – individuals were not color marked?
In our own recordings of colonial species, when several individuals were present, we made sure we were close enough to the focal bird and saw it opening its beak when singing. To eliminate any possible doubt we manually marked the recording when the focal individual was singing.

For the recordings collected from sound archives, we avoided attributing songs from multiple individuals to a single one by selecting only recordings where we were certain that only one individual was singing. These recordings are usually accompanied by descriptions from contributors specifying whether the recording features a single individual or multiple ones.
If there was no description, it was often easy to detect and avoid the presence of multiple individuals, as overlapping songs and differences in signal-to-noise ratios (due to varying distances between the microphone and each singer) provided clear cues.

In cases where there was neither a description nor an obvious identification of a focal individual from the spectrogram, only a single song (and thus a single individual) was extracted.
In all instances, if any doubt remained, we simply discarded the recording.

- Line 305-307: I am not sure I understand this sentence – maybe re-word?
We improved the purpose by explaining in more details: “In this model, the response variable is a measure of similarity between a pair of species (the complement of the Jaccard distance, measured using the presence-absence matrix of categories), and the predictors are categorical variables assessing how similar the pairs are in terms of mating system, habitat openness, and colony size. Phylogenetic relatedness, a measure of similarity between pairs of species (the complement of genetic distance), is usually included as a random factor. However, since the response variable here is also a distance measure between species, we included phylogenetic relatedness as a fixed factor” (l.315-322)

- Line 336-337: I suggest “received further statistical support…”
Done (l.351)

- Line 344-346: How robust is this approach? If the authors compare “solitary” vs “grouped” do they get similar findings? What about if they compare “large” vs “all other” groups?
Thank you for this interesting comment, that was also pointed out by the reviewer 1. Following your suggestion, we conducted two additional Phylogenetic Path Analyses to test the robustness of this analysis. One compares “solitary” vs. “all other” groups, and the other compares “large” vs. “all other” groups. The results of the three PPA (Figure 3) remain consistent despite the reorganization of the groups.

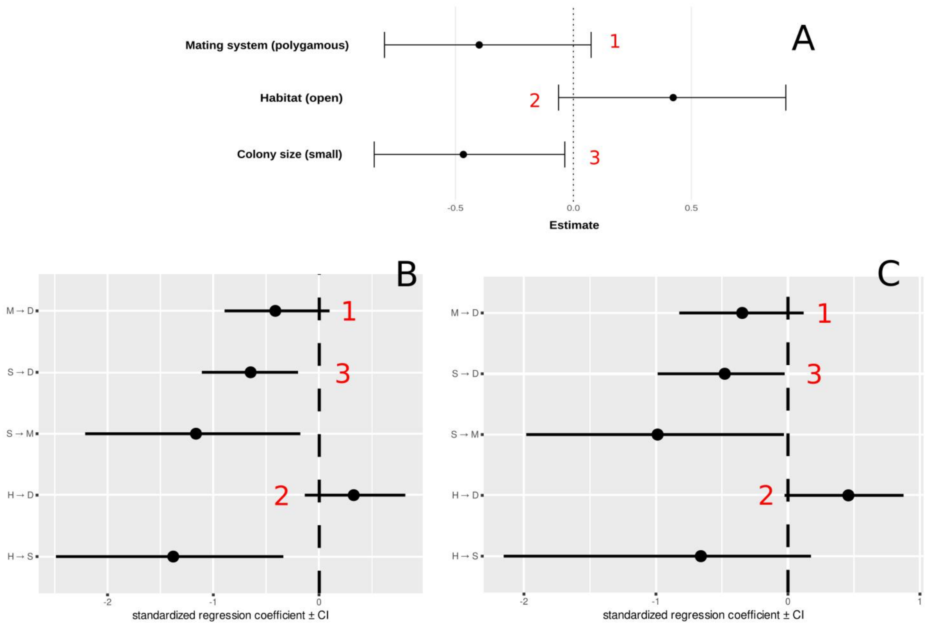

Figure 3: Comparison between (A) the PPA with 2 colony size categories ‘small’ and ‘large’, (B) the PPA with 2 categories of species ‘colonial’ vs ‘noncolonial’ and (C) the PPA with 2 categories of species ‘large’ vs ‘all other’. S-->D correspond to the effect of sociality on song diversity (3). M-->D correspond to the effect of mating system on song diversity (1) and H-->D the effect of habitat on song diversity (2). S-->M and H-->S corresponds to the effect of sociality on mating system and habitat on sociality.

- Line 357-358: I don’t have the reference on hand, but Robin Dunbar has a 2012 paper (in the Philosophical Transactions theme issue on social complexity and communication) on the importance of social bonding to vocal complexity – perhaps that discussion/idea is relevant to the finding here?
Thank you for this suggestion. The article you are referring to is probably “Bridging the bonding gap: the transition from primates to humans”. We now cite another article of Dunbar in the following sentences, to include this idea (l.519-527): “Finally, it is important to point out that mating system not only reflect the intensity of sexual selection but also a form of socialization. Indeed, monogamous species may exhibit stronger social bonds, where

members of a pair bond must have a much finer understanding of each other's needs and intentions (Dunbar & Shultz, 2010). For instance, a comparative analysis in primates showed that social bonds, measured by the time individuals spent grooming, were positively correlated with vocal repertoire size (McComb & Semple, 2005). Consequently, it is possible that this phenomenon explains why monogamous species have a greater diversity of songs than polygamous species."

- Line 373-374: I am not sure I understand the figure – particularly Figure 1b and its caption information. Do negative numbers indicate a negative effect – as in increased colony size is linked to a decrease in song diversity? That seems counter to the authors' interpretations and arguments.

As in Figure 2, the x-axis of Figure 1b represents the model's estimated parameters. For this analysis, colony size was recoded as a binary variable with two levels: "small" and "large". The "large" category is used as the reference group (and therefore does not appear in the forest plot). Consequently, the estimate of -0.47 represents the mean difference in song diversity between species living in small colonies compared to those in large colonies. If the model's link function were the identity function, the difference would be additive. For instance, if species in large colonies had an average song diversity of 6, the model would predict a mean diversity of 5.53 (6 - 0.47) for species in small colonies. However, in this case, the model uses a log link function, meaning the relationship is multiplicative rather than additive. As a result, the average song diversity of species living in small colonies is $e^{-0.47}$ times that of species in large colonies. This implies that species in small colonies have a song diversity that is approximately 0.64 times that of species in large colonies, or, conversely, species in large colonies exhibit a song diversity that is 36% higher, on average, than those in small colonies.

- Line 424-425: I suggest "proximity predicted both song diversity…" or "proximity was associated with both song diversity…"

Done (l.446-447)

- Line 492: I suggest "no data currently exist…"

Done (l.518)

- Line 495: Do the authors mean "interspecific" here rather than "intraspecific"?

Indeed, it was a mistake. We modified (l.529).

- Line 521-522: In comparative studies like this on the Social Complexity Hypothesis, it seems often difficult to determine whether social complexity might be driving vocal complexity (as the authors are arguing here) or whether vocal complexity that might evolve for other reasons might be making greater social complexity possible. I wondered if the phylogenetic signal detected here might actually be support for the latter interpretation. If the phylogenetic effect is considered 'strong', might this bolster an argument that increased song complexity (phylogenetically driven) is driving social complexity?

We completely agree with you that song complexity could influence social complexity. However, we don't believe that the presence of a phylogenetic signal predicts the direction of causality. Our results on the effect of phylogeny suggest that song diversity clusters within the phylogeny, even after accounting for colony size, habitat openness, and mating system. In other words, if all weaverbirds had the same colony size, lived in open (or closed) habitats, and were monogamous (or polygamous), closely related species would still have a tendency to share similar song diversity than distantly related ones. We don't see how this result would imply that it is song complexity that is driving social complexity in weaverbirds. That being said, your comment is very relevant, and to reflect this idea in the paper, we added the following sentence (l.469-472): "However, if a correlation was found, it does not determine the direction of causality, and a greater diversity of signals might lead to larger colony sizes, a hypothesis that could be further explored through ancestral state reconstruction."