Article

Current Biology

Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution

Highlights

- Online singing experiments enable large-scale music evolution studies
- Vocal, cognitive, and cultural factors cause oral transmission biases
- Transmission biases cause the emergence of diverse musical structures
- Social factors amplify or attenuate transmission biases

Authors

Manuel Anglada-Tort, Peter M.C. Harrison, Harin Lee, Nori Jacoby

Correspondence

manuel.anglada-tort@music.ox.ac.uk

In brief

Anglada-Tort et al. use online singing experiments to study oral transmission mechanisms in US and Indian participants. The results show how individual participant biases—vocal, cognitive, and cultural—shape the evolution of musical structures but that social biases are crucial for determining differences and similarities in resulting structures.

Anglada-Tort et al., 2023, Current Biology *33*, 1–15
April 24, 2023 © 2023 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cub.2023.02.070

CellPress

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

Current Biology

CellPress
OPEN ACCESS

Article

Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution

Manuel Anglada-Tort,[1,2,5,*] Peter M.C. Harrison,[1,3] Harin Lee,[1,4] and Nori Jacoby[1]
[1]Computational Auditory Perception Group, Max Planck Institute for Empirical Aesthetics, Grüneburgweg 14, Frankfurt am Main 60322, Germany
[2]Faculty of Music, University of Oxford, St Aldate's, Oxford OX1 1DB, UK
[3]Faculty of Music, University of Cambridge, 11 West Road, Cambridge CB3 9DP, UK
[4]Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, Leipzig 04103, Germany
[5]Lead contact
*Correspondence: manuel.anglada-tort@music.ox.ac.uk
https://doi.org/10.1016/j.cub.2023.02.070

SUMMARY

Speech and song have been transmitted orally for countless human generations, changing over time under the influence of biological, cognitive, and cultural pressures. Cross-cultural regularities and diversities in human song are thought to emerge from this transmission process, but testing how underlying mechanisms contribute to musical structures remains a key challenge. Here, we introduce an automatic online pipeline that streamlines large-scale cultural transmission experiments using a sophisticated and naturalistic modality: singing. We quantify the evolution of 3,424 melodies orally transmitted across 1,797 participants in the United States and India. This approach produces a high-resolution characterization of how oral transmission shapes melody, revealing the emergence of structures that are consistent with widespread musical features observed cross-culturally (small pitch sets, small pitch intervals, and arch-shaped melodic contours). We show how the emergence of these structures is constrained by individual biases in our participants—vocal constraints, working memory, and cultural exposure—which determine the size, shape, and complexity of evolving melodies. However, their ultimate effect on population-level structures depends on social dynamics taking place during cultural transmission. When participants recursively imitate their own productions (individual transmission), musical structures evolve slowly and heterogeneously, reflecting idiosyncratic musical biases. When participants instead imitate others' productions (social transmission), melodies rapidly shift toward homogeneous structures, reflecting shared structural biases that may underpin cross-cultural variation. These results provide the first quantitative characterization of the rich collection of biases that oral transmission imposes on music evolution, giving us a new understanding of how human song structures emerge via cultural transmission.

INTRODUCTION

Singing—the vocal production of musical sounds—is a unique feature of human culture, displaying an extraordinary diversity of forms cross-culturally while also sharing certain structural and functional properties.[1–4] Singing obeys similar acoustic and physiological principles to speech but nonetheless has distinctive features, such as the use of limited sets of stable pitches that make up musical scales.[5,6] This form of vocal production is a universal communicative modality for music, playing important social functions across cultures.[7,8] Thus, singing is a fascinating phenomenon for studying the biological and cultural foundations of music evolution.

For most of our evolutionary history, oral transmission has been the main mechanism by which songs are passed through generations. However, this simple act of transmission—hearing and singing back a song—does not result in the perfect transfer of information. Singers are likely to introduce some variation into their vocal productions, either accidentally or on purpose. Naively, one might expect that this variation is random, but in practice, oral transmission is thought to shape musical systems in non-random ways that reflect human transmission biases, such as those imposed by motor constraints and cognitive abilities.[9–13] Crucially, the outcome of oral transmission may depend not only just on biases of individual learners but also on the patterns of social interactions by which cultural transmission takes place, such as underlying social networks and population structures.[14,15] Although such processes of oral transmission likely played important roles throughout our evolutionary history,[16–18] explaining how they contribute to the structures of human song remains a key challenge for cognitive science.

It is possible to study oral transmission processes in controlled experiments via *iterated learning*, a powerful experimental

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

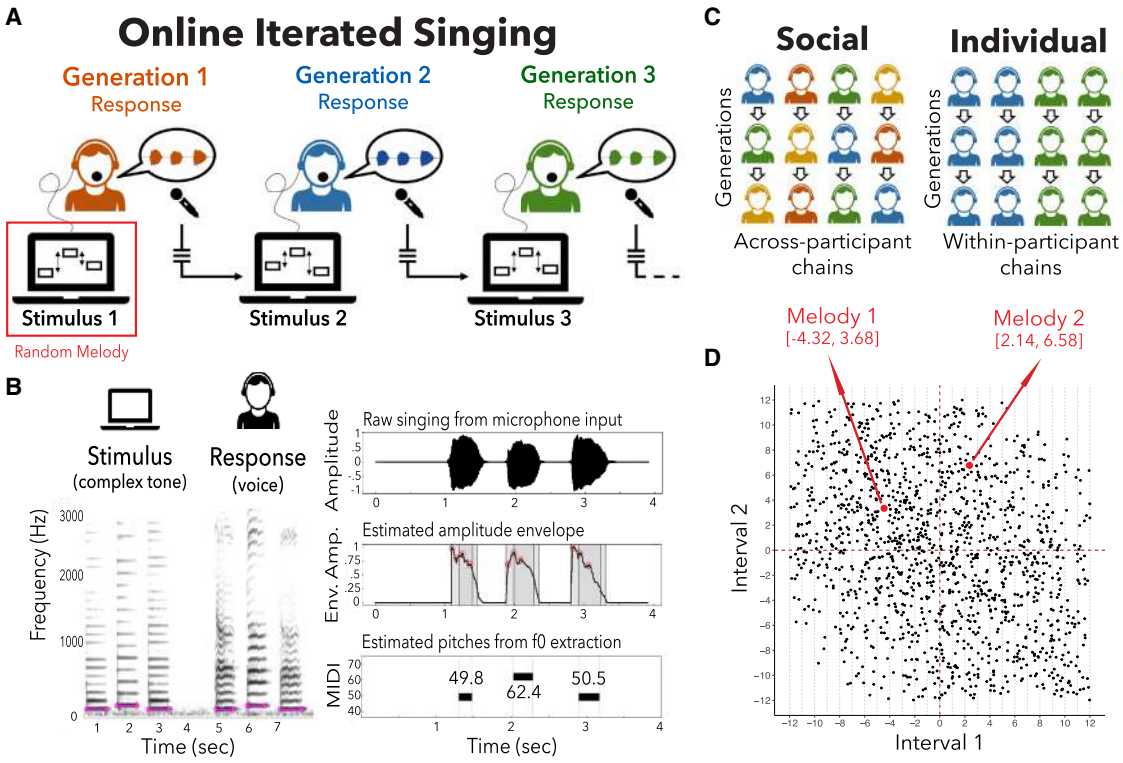

Figure 1. Online iterated singing paradigm

(A) Participants hear a sequence of tones generated by a computer and reproduce it by singing back. Vocal reproductions are automatically analyzed, synthesized, and played to the next participant as the input melody, using a pitch-roving technique to minimize inter-trial dependencies and adjust melodies to participants' singing range (see pitch roving procedure).

(B) On the left, a spectrogram of a three-tone melody and corresponding vocal reproduction (f0 indicated in pink). On the right, a schematic of the singing transcription procedure to estimate MIDI notes from the recording using f0 extraction techniques (see singing transcription technology).

(C) Melodies can be transmitted across participants (social transmission) or within participants (individual transmission).

(D) The entire stimulus space of three-tone melodies can be defined along two continuous dimensions, one for each interval in the melody (dots represent melodies). We initialize our experiments by randomly and continuously sampling melodies from this space (see melody generation). Melodies on the top-right and bottom-left corners are less likely due to the max pitch range parameter used to sample melody tones (see Table S2 for design parameters of all experiments). Melodies are represented using standard MIDI notation (see melody representation).

paradigm for studying the evolution of complex cultural phenomena such as language and technology—also known as *transmission chain experiments*.[19–22] Proof-of-concept iterated learning experiments have also been conducted for musical domains, including rhythm[23–25] and melody.[26–29] They have shown that iterated transmission yields the emergence of particular structures that can be interpreted in terms of participants' motor, cognitive, or cultural biases. However, iterated learning experiments in language and musical domains tend to suffer from a limited scale, only testing a few tens of participants and transmission chains at a time. This makes it particularly hard to systematically explore the vast space of evolutionary possibilities or to draw statistically reliable conclusions about underlying mechanisms.

Here, we introduce an automatic online pipeline that allows us to scale up cultural transmission experiments in complex production modalities by orders of magnitude. Our method is unique in leveraging online data collection while preserving a sophisticated and naturalistic task: singing. We focus in particular on the domain of *musical melodies*, short sequences of tones that make up the identity of songs. Participants are initially presented with a random sequence of tones and asked to reproduce it by singing (Figure 1A). Their reproductions are automatically synthesized in real time to generate new input melodies for the next participants (Figure 1B; see singing transcription technology). To simulate cultural transmission, we examine social transmission across participants, where chains are completed by different individuals. In some experiments, however, we also examine individual transmission, where the entire chain is completed just by one participant (Figure 1C). Importantly, our method does not assume culturally specific knowledge about musical scale systems *a priori*, such as the Western 12-tone chromatic scale. Instead, we randomly sample melodies from a continuous intervallic space (Figure 1D; see melody generation). Our method is fully automated and works efficiently over the internet using standard computers, massively increasing the reach, diversity, and scalability of data collection.

Mechanisms of melodic transmission

The oral transmission of melodies depends on several psychological and physical processes. First, a listener must "hear"

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

the melody, with their perception involving various auditory processes such as pitch estimation and interval extraction. The resulting mental representation of the melody must then be encoded in memory and retained for some period. At a later point, the listener must reconstruct the melody from memory and attempt to sing it. Their singing of the melody involves translating from a mental representation to a set of physical movements, most importantly of the diaphragm (for determining breath control) and the larynx (for determining pitch content).[5,6,30]

The different steps of oral transmission may each contribute particular biases to the transmission process. Melodic perception is naturally constrained by the frequency sensitivity of the ear, and the spectral/temporal pattern-matching processes responsible for pitch perception[31,32]; it is also heavily influenced by the listener's cultural background,[33,34] which causes the listener to interpret melodies in terms of prelearned schemata such as pitch categories (scales)[35] and temporal categories (rhythms),[23,36] as well as more dynamic expectations that develop over the course of the melody (melodic expectations).[37,38] These influences may be further magnified by memory encoding and retention, since culturally learned schemata provide an important scaffolding for the parsimonious retention of information in memory.[39,40] Finally, the vocal reproduction of a melody may induce further biases corresponding, for example, to the individual's limited vocal range or agility.[12]

These components may be responsible for inducing biases on a "local" level, from one transmitted melody to the next. However, cultural transmission requires learning information socially from others, either via imitation, teaching, or mere exposure.[41] How do population-level structures depend on the underlying dynamics of social interactions? One possibility is that all individuals share strong biases for music, constraining them to produce only certain musical structures over generations. Computational modeling of iterated learning[19] predicts that this is indeed the case when the stationary distribution of the model depends only on the learners' priors, which are assumed to be similar across individuals and unchanged during the transmission process. Another possibility is that individuals instead have weak and diverse musical biases, which are amplified or attenuated over time through multiple social interactions. This creates a shared structural compromise that largely depends on the underlying patterns of social interactions.[42–44] Previous studies on human and animal communication have tested these competing hypotheses by comparing iterated learning results in the presence and absence of social interactions.[23,44–46]

Although previous research has discussed how such processes of oral transmission might have contributed to the evolution of human song,[10,17,18] it has been thus far impractical to test which mechanisms are important in practice and what effects they end up having on musical structures. Here, we address this with a series of 12 behavioral experiments with 1,797 online participants (see STAR Methods for details on recruitment, demographics, and experiment implementations). Our results have three primary contributions. First, we show that our method efficiently characterizes the effects of oral transmission on melody, shaping initially random sounds into more structured systems that increasingly reuse and combine fewer elements (small pitch sets, small intervals, and arch-shaped contours). Second, we probe the relative contribution of underlying individual mechanisms through a series of carefully designed experimental comparisons that activate or bypass particular mechanisms—i.e., production constraints, working memory, and cultural exposure. Finally, we study the relationship between individual and social transmission biases by comparing music evolution in different participant populations (United States vs. India) and transmission chain designs, where participants either imitate their own (individual transmission) or other participants' productions (social transmission).

RESULTS

Oral transmission shapes the evolution of melodies via iterated singing

Short melodies

We begin by examining the effect of oral transmission on short melodies composed of three tones (or two intervals). We explored this space with 590 across-participant chains with 10 generations (5,900 singing trials) and a total of 188 US participants. Figure 2A shows the results of the iterated singing experiment across generations. Oral transmission shaped initially random tones into melodic structures that increasingly reused fewer interval combinations. Indeed, melodies in the last three generations of the experiment are concentrated around a few locations (Figure 2B), displaying a rich structure that resembles Western discrete scale systems. For example, a popular area in the space consists of arch-shaped melodies going up and down in pitch (bottom-right quadrant), mostly peaking in the perfect fifth (intervals [7, −7]). We also used a peak-finding algorithm to identify significant peaks in the marginal distribution of the two melodic intervals (see peak finding; marginals are plotted at the top and right of Figure 2B). Statistically significant peaks (indicated as red dots) further reveal the existence of interval categories consistent with the Western 12-tone scale. For example, we can see significant peaks near locations close to integer semitones, such as peaks at −4.85 [−5.08, −4.62] and 4.80 [4.59, 5.02] semitones in the first interval and peaks at −7.10 [−7.42, −6.77] and 4.19 [3.92, 4.46] semitones in the second interval, using 95% confidence intervals (CI).

Next, we looked at the effects of oral transmission on structural properties of melodies. Table S3 provides the summary statistics for all trend analyses conducted in this study, using linear regressions with 95% CI derived from bootstrapping (1,000 replicates; see trend analysis). First, to quantify the emergence of melodic structure, we computed the entropy of the distribution of intervals using Shannon's entropy (see interval entropy). Interval entropy decreased significantly over time (Figure 2C), suggesting an increase in melodic structure. Second, we found that melodies were biased toward a small vocabulary of intervals (Figure 2D), shown by a significant decrease in the mean number of detected peaks in the distribution of intervals over generations (see interval vocabulary size). Third, we observed that melodic intervals became significantly smaller over time (Figure 2E), as indicated by the mean absolute interval size. For example, the proportion of intervals larger than 7 semitones declined from 36.38% [32.66, 40.09] in the baseline to 11.90% [8.06, 15.74] in the last generation of the experiment (95% CI). Finally, we calculated a measure of copying error corresponding to the

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

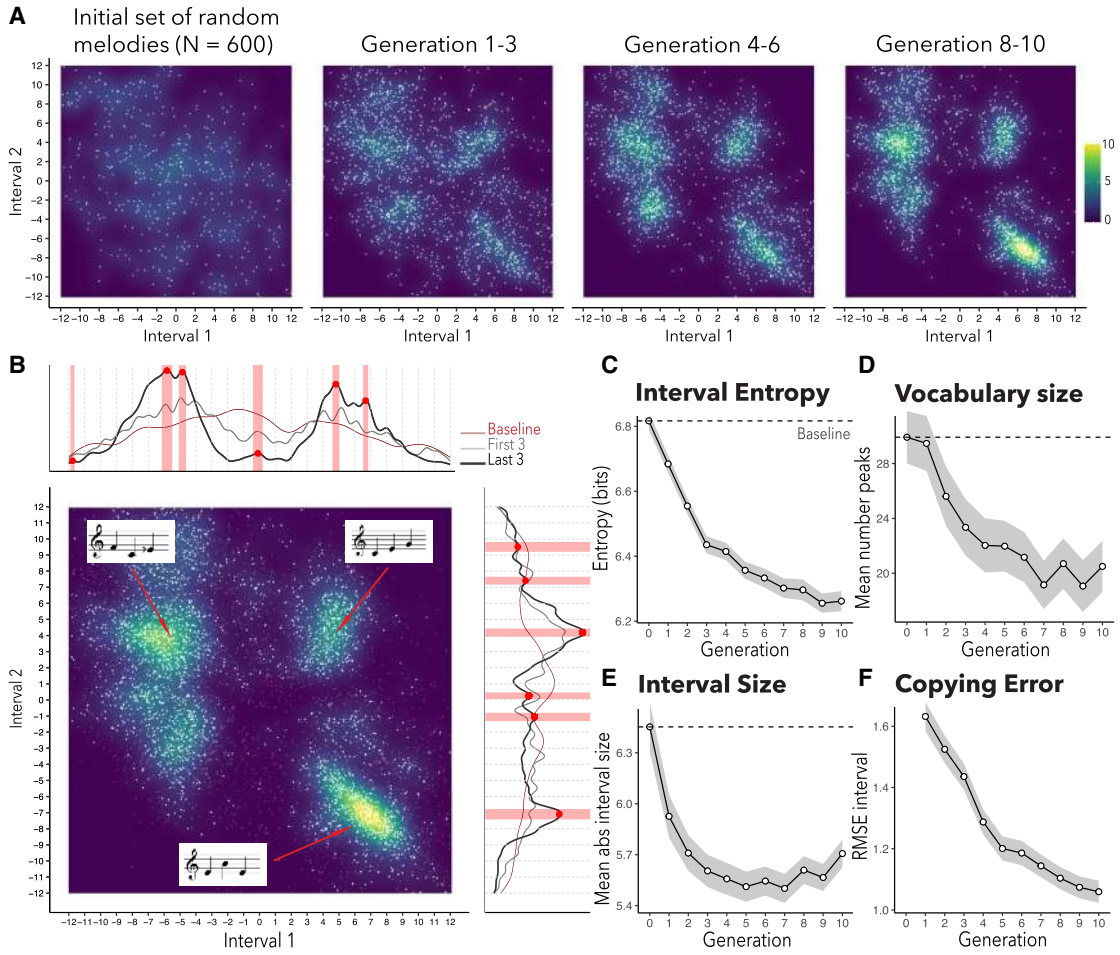

Figure 2. Oral transmission effects on short melodies

(A) The distribution of melodies over generations, using a two-dimensional kernel density estimate (see 1D and 2D Kernel Density Estimation (KDE)) over the locations of the melodies (density is expressed relative to a uniform distribution; yellow areas represent high density).

(B) KDE over the last three generations of the experiment. We plot the marginals of each interval separately on the top and right panels of the figure (black line), including the marginals of the random initial set of melodies (dark red line) and first three generations of the experiment (light gray line). Statistically significant peaks (see peak finding) are indicated by the red dots and shaded areas (95% CI).

(C–F) The effects of oral transmission on melodic features over generations (dashed lines indicate baseline values). Shaded areas correspond to ±1 standard error derived from bootstrapping (1,000 replicates). See Table S3 for summary statistics.

distance between the target melody and response (see copying error). Copying error decreased significantly over time (Figure 2F), suggesting that melodies became increasingly easier to learn and transmit. Overall, these findings are consistent with large-scale quantitative data showing that melodies cross-culturally tend to contain a small number of interval categories per octave (7 or less) and are composed of intervals of small size (less than 7 semitones).[1,4]

Long melodies

We also generalized our investigation to longer melodies. Experiment 2 studied five-tone melodies (159 chains, 51 US participants), whereas Experiment 3 studied variable-length melodies ranging from 4 to 12 tones (216 chains; 83 US participants), allowing melodies to change in their number of tones as they were passed across participants. Figure 3A shows the evolution of melodic features in the three singing experiments (statistics reported in Table S3). To keep the results comparable between experiments with different baseline levels, we linearly normalized the melodic features based on the baseline values at the start of the experiments (we use this strategy in all subsequent analyses; see comparing melodic features between experiments). As with short melodies, longer melodies exhibited a significant increase in melodic structure, manifesting as a significant decrease in the entropy of the distribution of intervals (see "interval entropy" in Figure 3A). Melodies also exhibited a bias toward a small vocabulary of intervals ("vocabulary size"; except for the variable-length experiment), and a bias toward small melodic intervals ("interval size"). Melodies in all experiments became increasingly easier to transmit over time ("copying error"), but this effect was larger in experiments with longer melodies, reflecting higher demands in the task. However, we also see differences in the evolution of melodic features depending on melody length. For

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

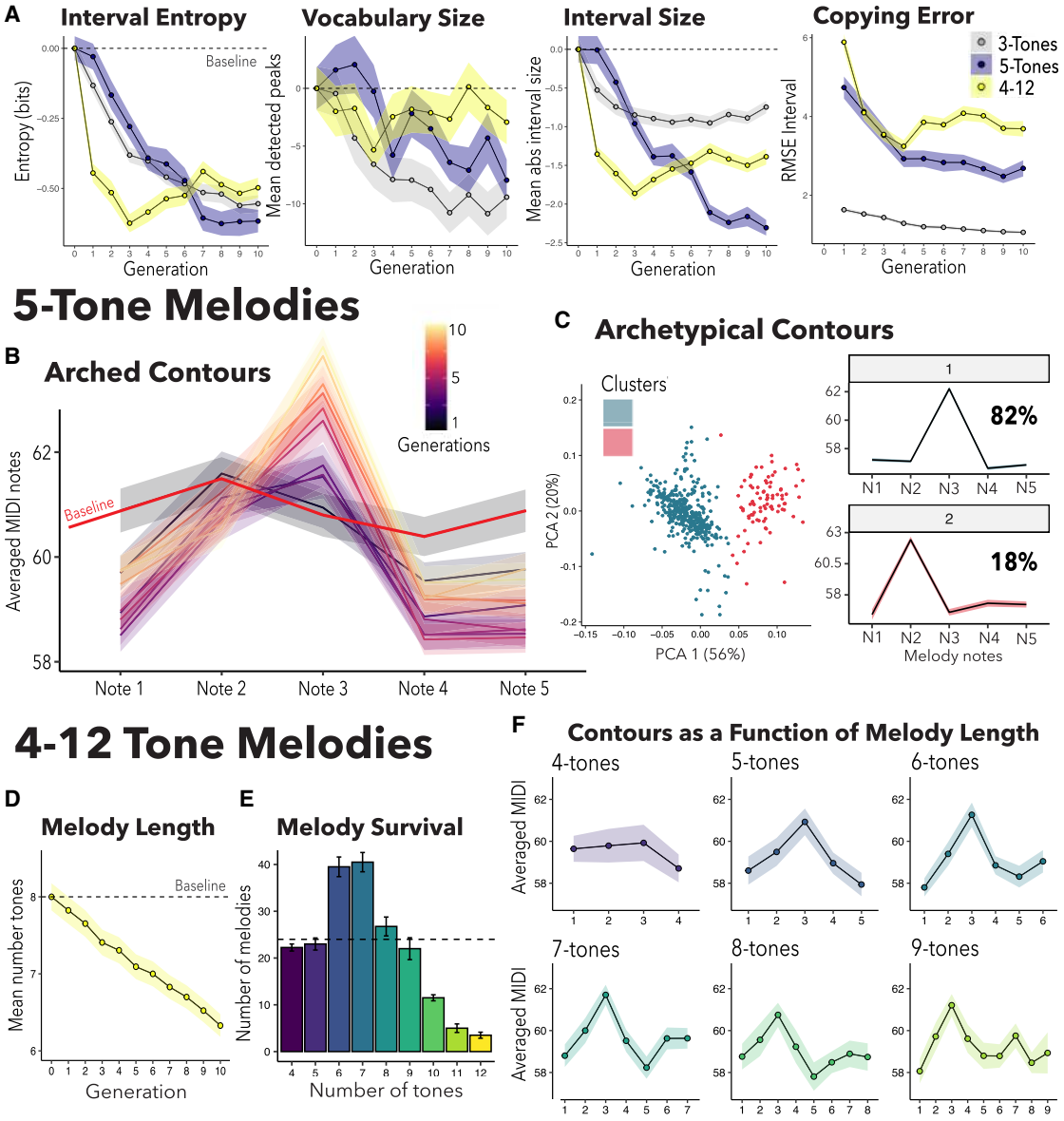

Figure 3. Oral transmission effects on long melodies

(A) The effects of oral transmission on melodies of different lengths (Experiment 1–3; see Table S3 for summary statistics). Melodic features are normalized based on baseline values (dashed line), except for copying error (see trend analysis). Shaded areas represent bootstrapped standard error (1,000 replicates).

(B) Evolution of melodic contour in Experiment 2 (error bars represent SE; see melodic contours).

(C) Clustering results (*k*-means) over the five melody tones in the last three generations. Left plot: melodies colored by cluster and projected over a two-dimensional PCA space (explained variance in brackets); right plot: average melodic contour in each cluster.

(D and E) (D) The average number of tones per melody over time (Experiment 3) and (E) the number of melodies of different lengths in the last generation of the experiment.

(F) Melodic contours in melodies of different lengths in the last three generations of Experiment 3 (error bars represent SE).

example, five-tone melodies exhibited the largest decrease in the mean absolute interval size, suggesting a higher tendency to use stepwise motion and small intervals.

Using data from Experiment 2 (five-tone melodies), we examined the evolution of melodic contours (the sequence of ups and downs in pitch), a key feature in melody cognition.[39] To visualize melodic contours, we calculated the mean MIDI value (and SE) for each tone in the melody across all melodies per generation (transposing all melodies to the same register; see melodic contours). As shown in Figure 3B, melodies evolved from flat melodic contours to distinctive arch-shaped contours. Indeed, a clustering analysis over the melody tones in the last three generations of the experiment (using *k*-means clustering over averaged and centered MIDI notes) revealed that most melodies eventually clustered around only two contour types, both consisting of minor variations of an arch-shaped contour (Figure 3C).

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

Finally, we looked at the effect of oral transmission on melody length (Experiment 3). The mean number of tones per melody decreased significantly over generations (Figure 3D). By the end of the experiment, the longest melodies (10–12 tones) almost disappeared, whereas melodies with 6 to 7 tones became the most popular (Figure 3E). We can also see that the emergence of arch-shaped musical contours was similar across melodies of different lengths, with the highest pitch near the third tone in the melody (Figure 3F). The prevalence of simple and arched musical contours has been found to be widespread both in bird[12] and human song across cultures.[1,4]

To summarize, we have demonstrated that oral transmission has profound effects on the emergence of melodic structures, shaping initially random tones into more structured systems that become increasingly easier to learn and transmit. Importantly, structural features emerging artificially from our experiments—small pitch sets, small pitch intervals, and arch-shaped melodic contours—are largely consistent with widespread musical features observed cross-culturally.[1,4]

Production constraints, working memory, and cultural exposure cause oral transmission biases

Musical structures emerging from our experiments can be seen as adaptations that arise from the transmission bottleneck imposed by individuals' capacity to process and produce music. In the following experiments, we present a series of experimental manipulations to probe underlying mechanisms at play and study their effects on melodic transmission. Specifically, we study the effect of vocal constraints, working memory, and cultural exposure.

Production constraints

We combined different experimental paradigms with computer simulations to explore the relative contribution of production and perception. First, we examined melodic transmission in the absence of vocal constraints by conducting an iterated learning experiment where participants copied melodies with a slider rather than their voice (Experiment 4: 369 chains and 327 US participants). We compared this experiment with a control singing experiment using a similar design (Experiment 5: 398 chains and 122 US participants). To keep the transmission process comparable between sliders and vocalizations, we transmitted stimuli composed of one melodic interval only (two tones played sequentially); in the slider experiment, we also implemented an aggregation technique to summarize the slider responses of multiple participants using the median[47] and a performance incentive (see iterated slider imitation). These techniques effectively reduced production noise in slider responses, which was comparatively higher than in vocal productions. The resulting stimulus space can be represented as a one-dimensional horizontal line, where each location corresponds to a unique interval. Participants were asked to match the target interval simply by moving a slider horizontally along the line (Figure 4A), keeping the first tone of the interval constant within each trial.

The aggregated results in the last generations of the two experiments reveal striking differences (Figure 4B). Melodic intervals in the singing experiment exhibited a relatively structured distribution (dashed black line), similar to the one obtained in Experiment 1. However, melodic transmission via sliders produced a comparatively less-structured distribution featuring a strong bias toward large intervals (see significant peaks at −12.07 [−12.38, −11.77] and 12.19 [11.96, 12.42] semitones). Figure 4C shows the evolution of melodic features in the two experiments (statistics reported in Table S3). Interval entropy decreased more readily when intervals were transmitted orally than with sliders. There was a similar but smaller trend in interval vocabulary size. However, the largest difference between the two experiments was the mean absolute interval size (see "interval size" in Figure 4C): melodies transmitted via sliders were significantly biased toward large intervals, whereas melodies transmitted orally were biased toward small intervals. Finally, copying error indicated that imitating melodies with sliders was significantly harder than with the voice, but in both conditions, melodies became significantly easier to transmit over time.

These results show that vocal constraints strongly facilitate melodic evolution, speeding the emergence of structural features such as vocabulary reuse and small interval sizes. By biasing transmission toward small intervals, vocal constraints restrict the available stimulus space and enable it to be explored more efficiently. One possible explanation is that large intervals require sudden contraction/relaxation in the muscle controlling vocal fold tension[12] and thus are harder to produce by the vocal system than smaller ones. To explore this kind of vocal constraint, we conducted a simulation experiment to model how melodies would evolve if oral transmission was shaped only by a simple (polynomial) function based on the interval size and direction, with additional independent stochastic production noise (see experimental simulations). The results of this simple model can account for important features observed in human data (Figure 4G; see Figure S2 for an example of the model across generations), such as the bimodal distribution of intervals with two major modes and a relatively large dip in between. However, this model fails to capture more nuanced features, such as the emergence of certain peaks observed in the singing data.

One limitation of the slider experiment is that it introduces new kinds of production biases—i.e., controlling and manipulating a slider. To isolate the effects of perceptual biases and study how they may relate to singing data, we ran a subjective preference experiment that did not contain any production component (Experiment 6). Specifically, we used a dense rating paradigm[48] to construct a detailed map of subjective preferences for melodic intervals, derived from 15,000 stimuli sampled randomly and uniformly from the range [−15, 15] semitones (see dense rating paradigm). A total of 415 US participants contributed to the experiment by listening and rating intervals using a "pleasantness" scale (Figure 4D). The aggregated ratings provide a highly detailed characterization of perceived melodic pleasantness in Western music (Figure 4E). Despite sampling intervals continuously from the space, we see clear peaks around integer locations that characterize the Western 12-tone chromatic scale, such as the octave (peaking at 12.09 [11.97, 12.21]), perfect fifth (peaking at −7.06 [−7.17, −6.95] and 7.10 [6.90, 7.29]), and perfect fourth (peaking at = −5.05 [−5.22 −4.86] and 5.10 [4.89, 5.31]).

We examined how well the intervals produced in the singing experiments aligned with this melodic pleasantness profile.

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

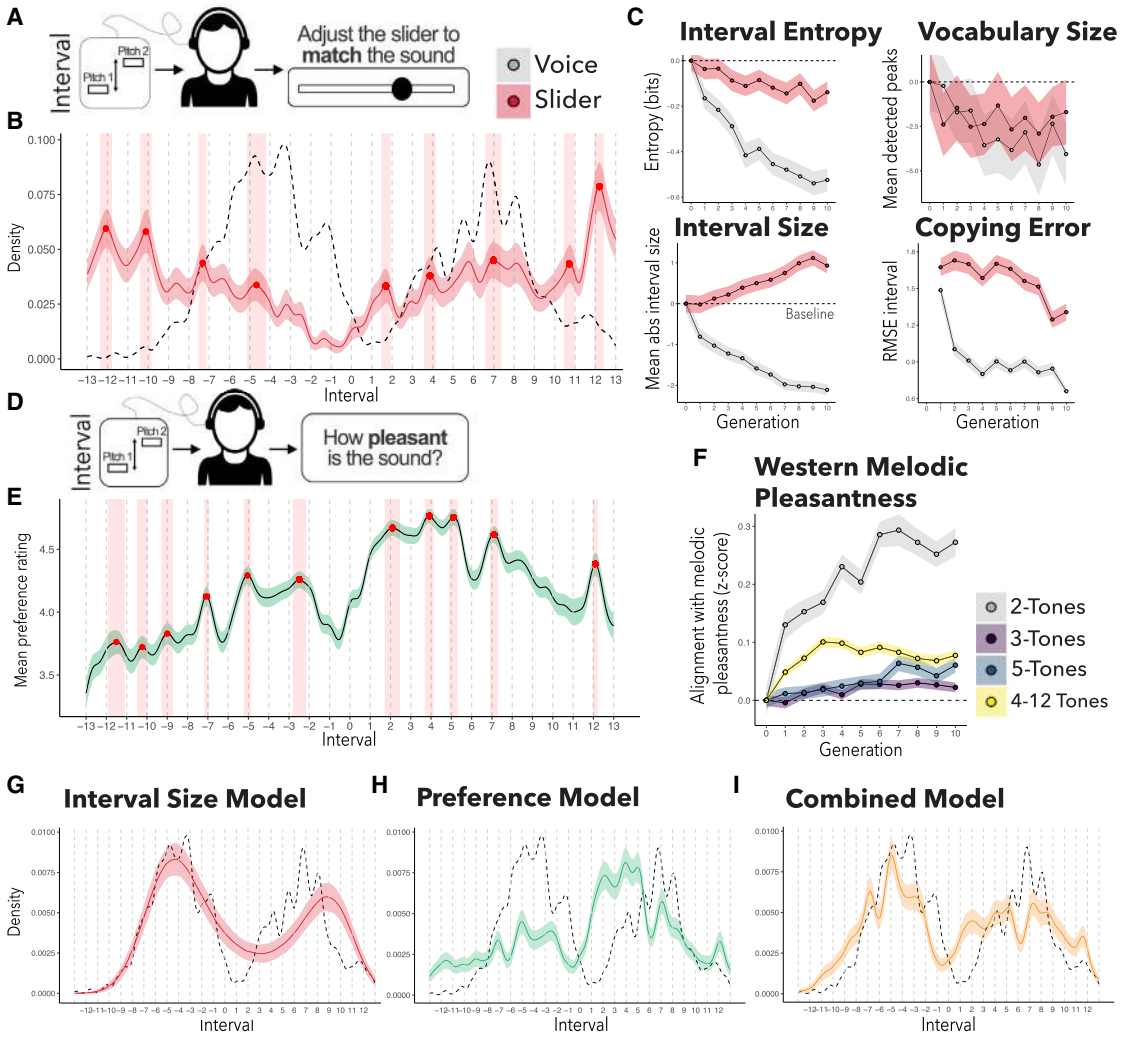

Figure 4. Production constraints

(A) Participants heard melodic intervals and were asked to match them using a slider (Experiment 4).

(B) The distribution of intervals in the last three generations of the slider (red line; Experiment 4) and control singing experiment (dashed black line; Experiment 5).

(C) The effects of oral transmission on melodic features (statistics reported in Table S3). Melodic features are normalized based on baseline values (dashed line), except for copying error (see trend analysis).

(D) Participants heard melodic intervals and were asked to rate their pleasantness (Experiment 6).

(E) The aggregated results of the rating experiment provide a highly detailed characterization of melodic pleasantness in Western music. Statistically significant peaks are indicated by the red dots and shaded areas (95% CI; see peak finding).

(F) Melodies in the singing experiments became increasingly more aligned with Western melodic pleasantness.

(G–I) Results in the last three generations of the simulation models based on (G) interval size and direction, (H) subjective preferences, and (I) a combined model (see experimental simulations; see Figure 2S for data across generations). Shaded areas in all plots correspond to ±1 standard error derived from bootstrapping (1,000 replicates).

The results indicated that orally transmitted melodies in all experiments became increasingly more aligned with Western melodic pleasantness (Figure 4F). We then conducted a simulation experiment to see how melodies would evolve if oral transmission was shaped solely by subjective pleasantness (see experimental simulations). This simulation translates preferences (a subjective utility function) to a perceptual "prior" and then uses a Bayesian serial reproduction model to predict participants' responses.[47,49,50] The results of this simulation can account for some features of the data (see Figure 4H; see Figure S2 for example data across generations), such as the emergence of some interval categories (e.g., peaks at −5, 4, and 7 semitones) and avoidance of others (e.g., the tritone). However, the preference-based model is insufficient for capturing the structure obtained from human singing.

Together, these results demonstrate that vocal constraints are necessary to converge to melodic structures that characterize human song. However, to account for more nuanced features, it is necessary to also consider perceptual biases (e.g., melodic pleasantness) and cognitive biases (e.g., learned schemata). Finally, we explored a model combining

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

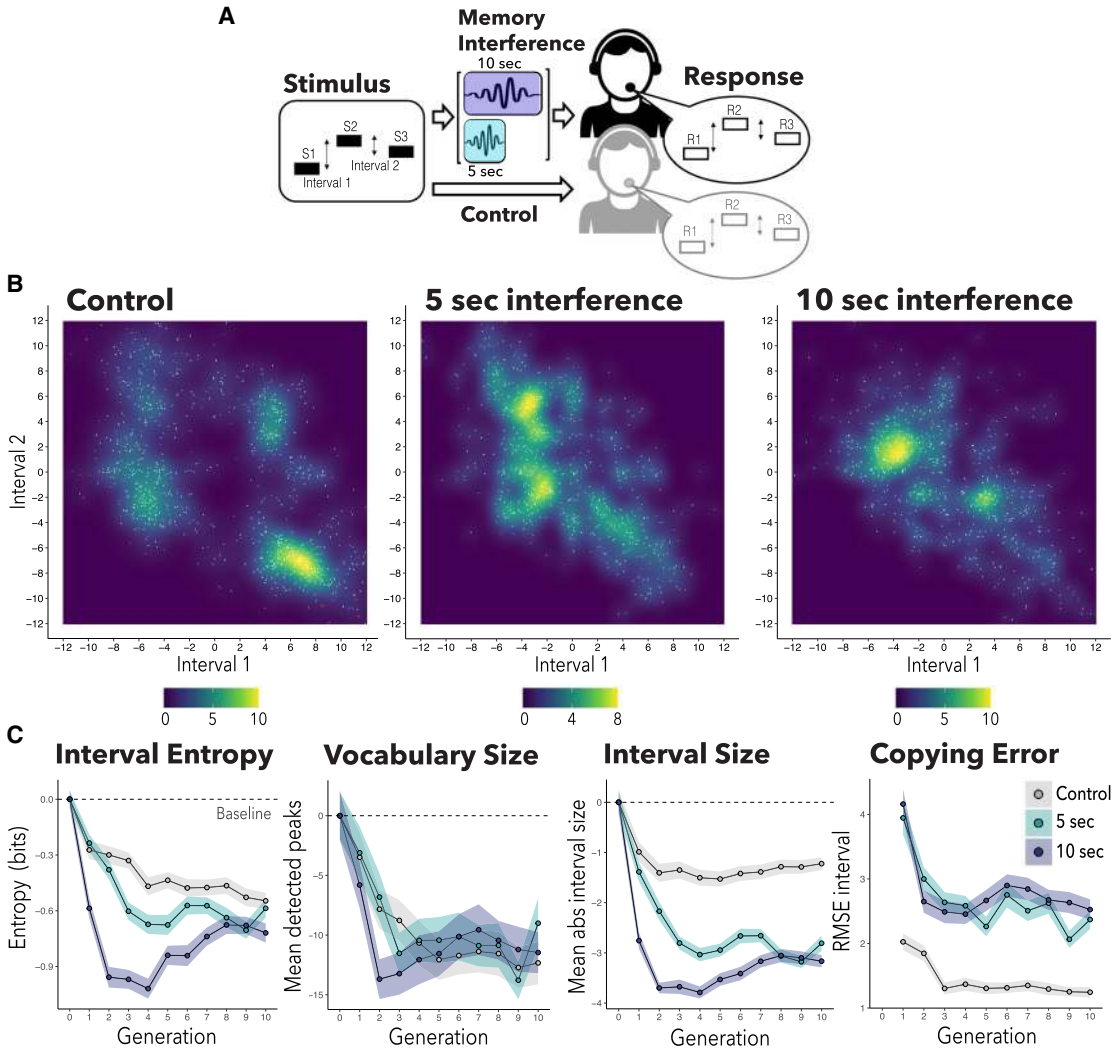

Figure 5. Working memory

(A) In the memory interference conditions (Experiment 7 and 8), participants heard three-tone melodies, followed by a random sequence of tones played at fast tempo (auditory interference) either for 5 or 10 s. In the control condition (Experiment 9), participants performed the same procedure but without any interference.

(B) The distribution of melodies produced in the last three generations of the three experiments (see 1D and 2D Kernel Density Estimation (KDE); see Figure S3 for results across generations).

(C) The effects of oral transmission on melodic features (statistics reported in Table S3). Melodic features are normalized based on baseline values (dashed line), except for copying error (see trend analysis). Shaded areas correspond to ±1 standard error derived from bootstrapping (1,000 replicates).

both the interval size and preference models (see Figure 4I; see experimental simulations) and found that it captures important features from both models but still does not fully predict the empirical data.

Working memory

To examine the effect of working memory, we conducted iterated singing experiments manipulating the memory interference between the target melody and the singing response (Figure 5A). The magnitude of the memory interference was controlled by playing an auditory distractor stimulus after the target melody either for 5 s (Experiment 7: 240 chains and 100 US participants) or 10 s (Experiment 8: 240 across-participant chains; 105 US participants). The auditory distractor consisted of a sequence of random tones with an overall duration of 100 ms played at 250 ms inter-onset interval. We compared the results of these experiments with a control experiment with no additional memory interference (Experiment 9: 240 chains; 95 US participants), everything else being equal.

Figure 5B shows the aggregated results in the last three generations of the three experiments (see Figure S3 for results across generations). Melodic structure emerged in all conditions, but the size and exact distribution of the resulting melodies largely depended on the memory manipulation. We see that the larger the constraints imposed on working memory, the smaller the intervals and the simpler the melodic structures. Indeed, interval entropy indicated that melodic structure emerged most rapidly in the high-interference condition (10 s),

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

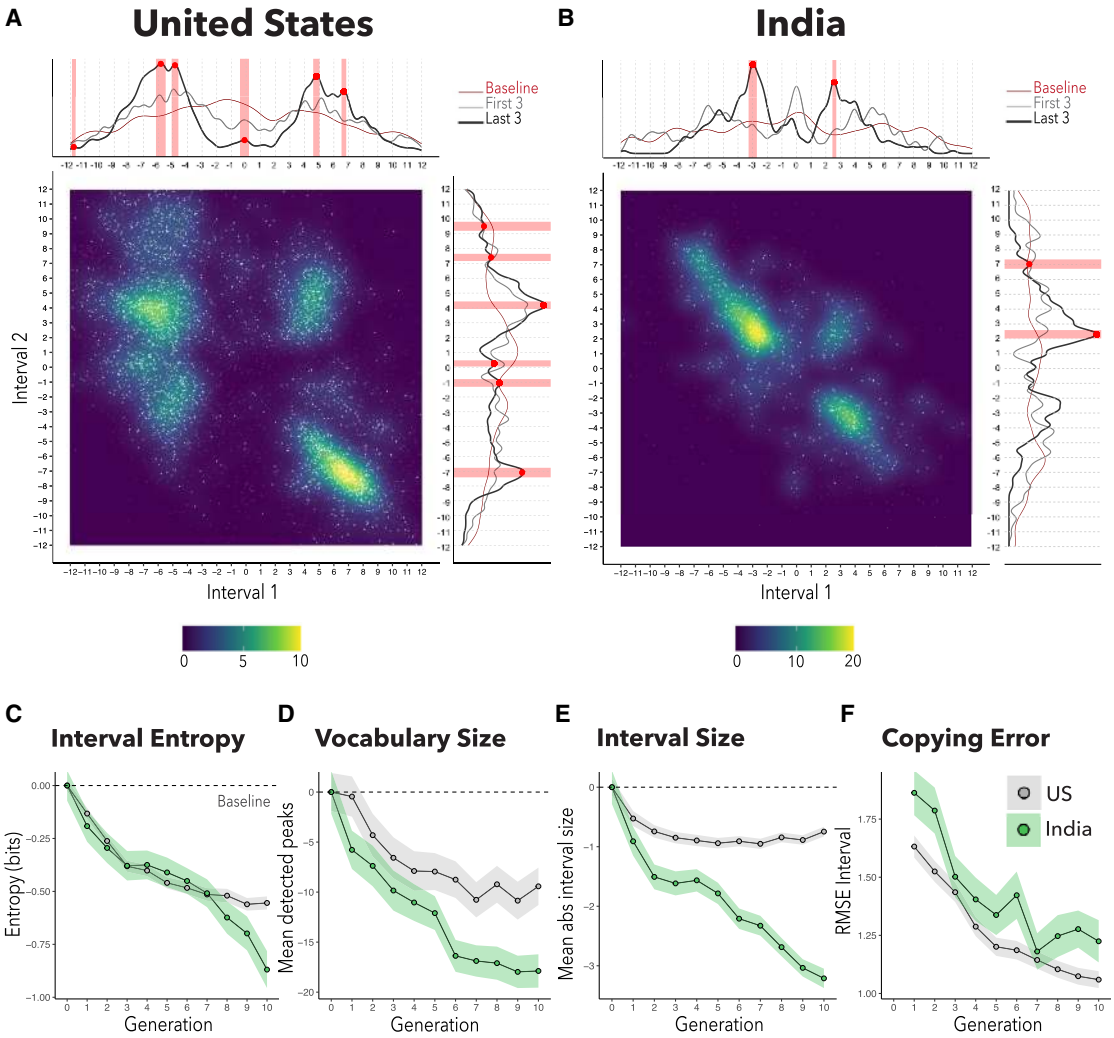

Figure 6. Cultural Exposure

(A and B) (A) The distribution of melodies produced in the last three generations of the singing experiments with participants recruited from the US (Experiment 1) and (B) India (Experiment 10; see 1D and 2D Kernel Density Estimation (KDE)). We plot the marginals of each interval separately on the top and right panels of the figure. Statistically significant peaks (see peak finding) are indicated by the red dots and shaded areas (95% CI).

(C–F) The effects of oral transmission on melodic features (statistics reported in Table S3). Melodic features are normalized based on baseline values (dashed line), except for copying error (see trend analysis). Shaded areas correspond to ±1 standard error derived from bootstrapping (1,000 replicates).

followed by medium interference (5 s) and no interference, although they all converged to similar levels by the end of the experiment (see "interval entropy" in Figure 5C). Interval vocabulary size decreased similarly in the three experiments ("vocabulary size"), whereas the mean absolute interval size showed a strong effect of memory: interval size decreased more drastically in the two memory interference conditions than in the control experiment, reaching an average size of about 2 semitones less (''interval size''). Finally, copying error reflected the difficulty of the experiments but improved significantly in all conditions ("copying error"; see Table S3 for the statistics).

These results demonstrate that memory constraints are an important bottleneck for evolution by oral transmission. Just 5 s of memory interference caused a substantial shift toward melodies with smaller intervals and simpler structures; the increased duration of memory interference further accentuated this effect.

Cultural exposure

The last individual transmission bias we explored concerns cultural exposure. We replicated the main iterated singing experiment (Experiment 1; US participants) using an online cohort of Indian participants (Experiment 10; 120 chains and 54 participants). A singing performance test conducted before the main task (see singing performance test) ensured that participants in the two groups were similar in singing accuracy (India: $M = 0.77$, $SD = 0.32$; US: $M = 0.64$, $SD = 0.30$, in semitones). Participants were also similar in demographic information and levels of musical expertise (see Table S1).

Figures 6A and 6B show the aggregated results in the last three generations of the two experiments (see Figure S4 for results across generations). The main effect of oral transmission

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

clearly replicated with the new group of participants: melodies evolved systematically toward more structured distributions. However, there were also interesting divergences between groups. Specifically, interval entropy decreased similarly in the two groups, but it was comparatively lower in the last three generations of the Indian group (Figure 6C) . Both groups were biased toward a small vocabulary of intervals (Figure 6D) and intervals of small size (Figure 6E), but these effects became more pronounced in the Indian group over time. Copying error indicated that both groups were comparable in terms of their improvement in meldoic transmissibility over time (Figure 6F; see Table S3 for statistics).

The largest difference between groups was found in the mean absolute interval size: melodic intervals produced by Indian participants were significantly smaller (3.25 [2.93, 3.57] semitones on average, 95% CI) than those produced by US participants (5.71 [5.55, 5.86] semitones). This finding is consistent with previous corpus studies showing that melodic intervals in South Indian melodies are smaller in size than that in Western Melodies.[51] This may be explained by the existence of different biases across participant groups favoring melodic intervals of different sizes. For example, US participants were biased toward "leap" intervals, including significant peaks around the perfect fifth (7 semitones) and perfect fourth (±5 semitones), which correspond to prototypical tonal intervals in Western music. By contrast, Indian participants were biased toward smaller intervals, including the major second (2 semitones) and minor third (3 semitones), both of which are key components of common Indian musical scales (e.g., pentatonic, hexatonic). These results clearly demonstrate cultural differences in the oral transmission that are not simply due to differences in singing abilities, presumably reflecting lifetime differences in musical exposure.

Social interactions modulate oral transmission biases

We identified several individual transmission biases underlying the effects of oral transmission on music evolution. We next asked whether the ultimate effect of individual biases depends on the dynamics of social transmission. We first compared the results of Experiment 1 (social transmission) with a new experiment using individual rather than social transmission (Experiment 11: 615 chains and 184 US participants). In individual transmission, each chain is completed only by one participant, measuring melodic transmission in the absence of social interactions (Figure 1C). To minimize memory effects, participants completed 4 full chains in parallel, allowing us to intersperse trials from different chains (see transmission chain designs).

Figure 7A shows the aggregated results in the last three generations of the two experiments using social and individual tranmision in the United States. The transmission chain design had profound effects on the outcome distribution of melodies: musical structures exhibited significantly higher diversity and less structure in individual rather than social transmission (see Figure S4 for results across generations). Indeed, interval entropy shows that melodic structure emerged more readily in social rather than individual transmission (Figure 7C; similar trends occurred for interval vocabulary size and average interval size; statistics are reported in Table S3). However, copying error decreased more drastically (and was overall smaller) when participants copied their own rather than others' productions (Figure 7C), suggesting that melodies were generally harder to learn and transmit during social transmission. Overall, these results show that the outcome of oral transmission largely depends on how melodies are transmitted across generations (socially vs. individually).

One possibility is that individual transmission created strong contextual effects, giving opportunities for participants to learn and evaluate their own productions (self-learning). However, we carefully designed our experiments to minimize contextual effects as much as possible, interspersing trials from multiple chains in parallel and randomly transposing melody tones in each trial. Another possibility is that the isolated nature of individual transmission preserved individual idiosyncratic biases over generations, causing slower convergence to melodic structures and higher diversity. If true, individual transmission may be a more effective method for uncovering granular divergences in musical biases than social transmission.[23] By contrast, social transmissions may speed up the emergence of population-level structures because participants are exposed to variations introduced by others, canceling out idiosyncratic biases that are not shared by all participants.

To explore this, we repeated the individual transmission experiment with a new group of online participants from India (Experiment 12; 223 chains and 73 participants). The results replicated cross-culturally (Figure 7B): melodies transmitted within participants exhibited significantly more diversity and less structure than melodies transmitted across participants (see Figures 7D and S5 for replication results in all melodic features; statistics reported in Table S3).

To directly compare the results in all conditions (social and individual transmission in the US and India), we plot the joint distribution of melodic intervals in Figures 7E–7H. Individual transmission provided a more granular characterization of musical biases in the two groups, shown by the higher number of significant peaks. In the US (Figure 7E), these peaks tended to fall around integer semitone categories that are largely consistent with the Western 12-tone chromatic scale, including the octave (peaking at −11.92 [−12.16, −11.69] and 12.19 [11.86, 12.52]) and perfect fifth (peaking at -7.01 [−7.13, −6.89] and 6.87 [6.4, 7.11]). Although some of these peaks were present in Indian melodies as well (Figure 7F), the results highlighted important cross-cultural differences. For example, there was an asymmetry in the Indian data, whereby major seconds (2 semitones) were rare in ascent but common in descent (peaking at −2.17 [−2.52, −1.81]. This may reflect aspects of Indian musical practice, where certain scales (often pentatonic or hexatonic) include ascending leaps of thirds that are filled in with stepwise motion when the scale descends. Indeed, the largest peak observed in the Indian dataset corresponds to an ascending minor third peaking at 3.25 [2.92, 3.58] semitones.

This cross-cultural comparison provides a particularly intriguing result: cross-cultural differences between the two groups were larger in social rather than individual transmission. Jensen-Shannon divergence (JSD), a measure of similarity between two probability distributions (see comparing distributions), indicated that the difference between distributions was

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

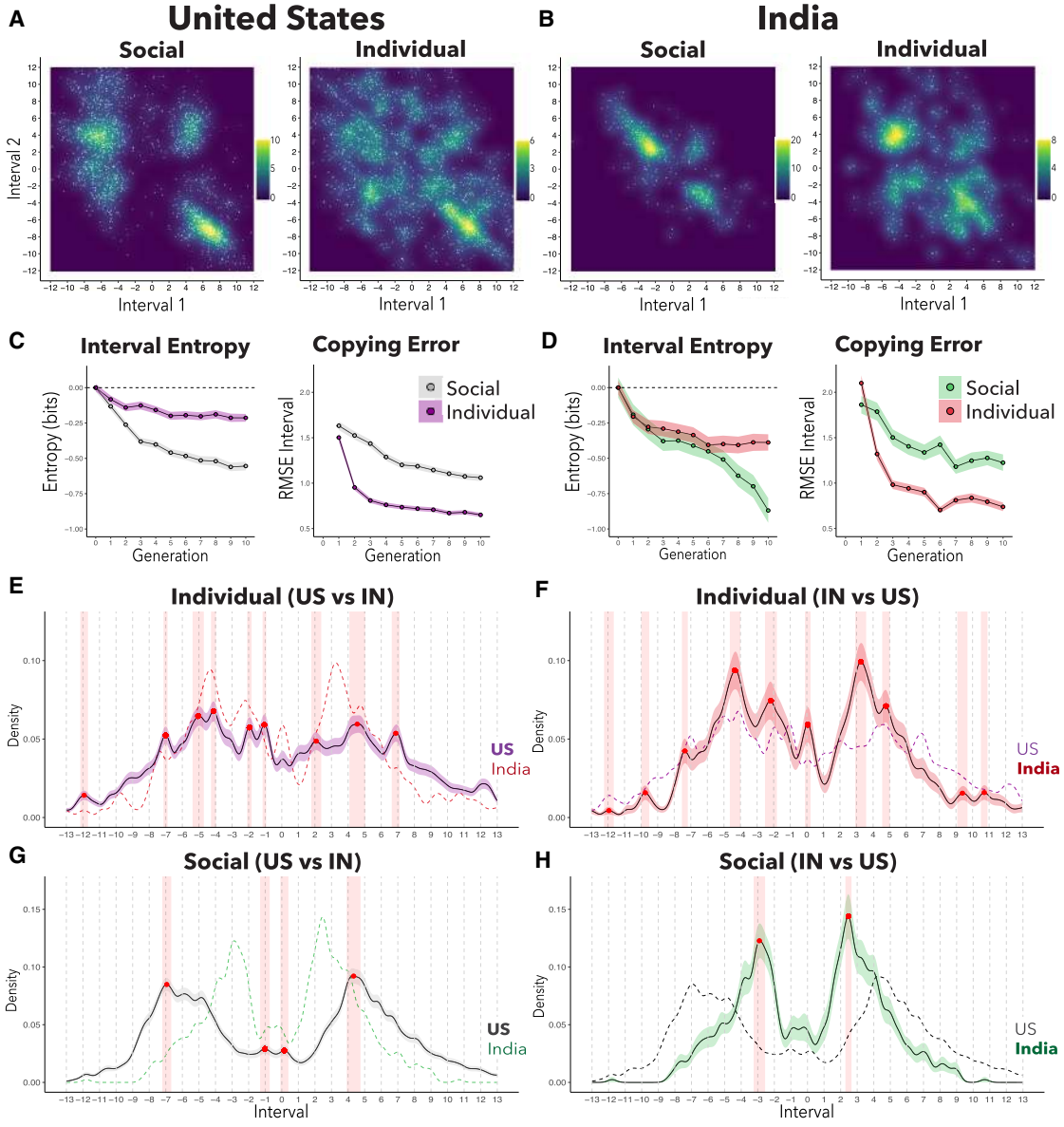

Figure 7. Social and individual transmission in the United States and India

(A and B) The distribution of melodies produced in the last three generations of the singing experiments during social and individual transmission in the United States and India (Experiments 1 and 10–12; see Figure S4 for results across generations; see 1D and 2D Kernel Density Estimation (KDE)).

(C and D) The effects of oral transmission on interval entropy and copying error (see Figure S5 for trends in all melodic features; see Table S3 for statistics). Interval entropy is normalized based on baseline values, whereas copying error is shown in absolute terms (see trend analysis).

(E–H) (E and F) The joint marginals of melodic intervals in the last three generations of the individual and (G and H) social transmission conditions in the United States and India (see Figure S6 for results comparing US participants with varying levels of musical expertise). Statistically significant peaks (see peak finding) are indicated by the red dots and shaded areas (95% CI). Shaded areas in all plots correspond to ±1 standard error derived from bootstrapping (1,000 replicates).

statistically larger in social transmission (*JSD* = 0.17 [0.13, 0.22], 95% CI) rather than individual transmission (*JSD* = 0.05 [0.03, 0.06]). We also observed larger cross-cultural differences in melodic features during social transmission (see Figure S5 for trends in melodic features comparing the two groups). This social attractor effect is visually apparent in Figures 7G and 7H, where social transmission caused a substantial shift toward different attractors in the two groups (the two major peaks in the US fell around −6.94 [−7.21, −6.67] and 4.33 [3.92, 4.75] semitones, whereas the two major peaks in India fell around -2.90 [-3.25, -2.56] and 2.47 [2.29, 2.65]). Interestingly, the overall topological distribution of melodies obtained in social transmission is remarkably similar in the two groups, featuring two prominent peaks with a dip in between.

Together, these results provide a clear cross-cultural replication of the effects of social transmission on melodic evolution: musical structures emerge faster and are more homogeneous when individuals copy others' productions rather than their

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

own. However, we also found that social transmission produced larger cross-cultural differences, suggesting that shared structural biases resulting from social interactions facilitate the emergence of cross-cultural differences.

DISCUSSION

We introduced an automatic online pipeline to perform large-scale cultural transmission experiments in the singing modality. Our results provide a highly detailed characterization of how mechanisms underlying oral transmission contribute to the emergence of cross-cultural similarities and differences in human song. The most salient feature of our results is that oral transmission shapes initially random sounds into structured systems. Over generations, participants introduced errors in their efforts to replicate the melodies they heard, giving rise to melodic features that were eventually easier to learn and transmit. Specifically, melodies were biased toward (1) a small vocabulary of intervals, (2) short lengths (5–6 tones per melody), (3) small interval sizes (less than a perfect fifth), and (4) arch-shaped melodic contours. These features are largely consistent with melodic features found in most musical traditions across the world.[1,2,4] However, our results also revealed deep differences in emerging structures across experiments and participant groups.

How can we explain the emergence of such structural similarities and differences? We found that, at a minimum, the outcome of oral transmission depends on a compromise between the biases of individual learners—vocal constraints, working memory, and cultural exposure—and the process of social transmission. Individual biases were the bottleneck of music evolution by oral transmission, determining the size, shape, and complexity of transmitted structures. For example, melodic features that were difficult to produce (Experiments 4–6) or remember (Experiments 7–9) were consistently less likely to survive the transmission process. However, the ultimate effect of individual biases on population-level structures depended on the dynamics of social interactions taking place during social transmission. When participants imitated their own vocal productions (individual transmission), musical structures converged slowly and were more diverse, reflecting idiosyncratic musical biases. When participants instead imitated other participants' productions (social transmission), musical structures emerged rapidly and homogeneously, reflecting shared structural biases.

We replicated these findings in the United States and India and found larger cross-cultural differences across groups during social rather than individual transmission. These results are surprising because they suggest that (1) population-level structures in vocal music depend on the underlying dynamics of social interactions and (2) shared structural biases resulting from social transmission can explain the emergence of cross-cultural differences in human song.

Limitations and future directions

Melodies in our experiments were constrained in several ways: they used discrete single-pitch tones, were fixed in rhythm, and in most experiments were limited to short sequences of three tones only. It is likely that additional mechanisms govern the transmission of more complex melodies, such as rhythm priors[23,36] or tonal expectations.[35] The use of discrete single pitches is also problematic because many musical traditions use non-discrete ornaments and melodic gestures, such as South Indian performance music.[52] Moreover, we used a pitch-roving technique to minimize inter-trial dependencies and adjust melodies to the participants' singing range (see pitch roving procedure). This technique may further limit the ecological validity of our experiments. Nonetheless, our paradigm can be easily extended to incorporate more complex musical elements, such as rhythm or longer melodies (see Experiments 2 and 3). Another promising extension consists of *naturalistic singing*, where instead of synthesizing melodies online, one could transmit participants' raw singing recordings. This will allow the study of pitch, rhythm, and dynamic variations within a single paradigm while also combining participants with different singing registers. Finally, our paradigm could be extended to track continuous human vocalizations, such as those produced by speech. This will enable exciting research into the intersection between speech and song, two cultural systems that have evolved through oral transmission.[5,16]

A priori, one might expect that peaks (frequent melodic intervals) obtained in our experiments should map directly onto the peaks in perceptual preferences and onto the prototypical intervals from relevant musical styles. Our results showed some overlap here (see peaks in Figure 7E), but they also revealed important differences. For example, in a purely perceptual task (see Figure 4E), we saw clear preferences for octaves and avoidance of tritones, but these results were less clear in the singing experiments (see Figures 7E and 7G). One potential explanation is production noise, which will make peaks and troughs less clear, especially when the interval is hard to produce. A second factor is production bias (e.g., interval compression), which may produce an undesirable interval (e.g., the tritone) from an attempt to produce a desirable one (e.g., the perfect fifth). A third relevant factor is melody length. With short melodies, it is harder to produce a clear tonality without using large intervals (e.g., outlining a major triad), and hence, cognitive biases toward tonal melodies may indirectly induce biases toward large intervals. Indeed, we found that shorter melodies tended to use larger intervals than longer melodies (see mean interval size between melody length conditions in Figure 3A). We see great potential in exploring further the interaction between tonality and melody length both in iterated singing and music corpus studies. However, any comparison with corpus data should control for melody length, as most corpora comprise long melodies primarily featuring small intervals.[51,53]

One should also take caution with the applicability of our results to explaining historical cultural processes. Naturally, culture is a complex and large-scale phenomenon, taking place in populations over multiple generations. Here, we have instead measured transmission events in a very short temporal scale and highly restricted interactions between modern humans (e.g., one-directional linear chains, absence of social context and feedback). However, the goal of our work was not to replicate the evolutionary history of music but rather to enable the study of oral transmission mechanisms in a controlled experimental setting. Iterated learning experiments have proven particularly useful to probe complex cultural transmission processes that would be otherwise hidden or very hard to infer from historical records.[20,21] We are excited about the potential

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

of incorporating more complex transmission dynamics within our iterated singing pipeline, studying for example the role of popularity dynamics,[54] selection biases,[55] or social network structures.[14,15]

Finally, we studied oral transmission processes across different groups of participants recruited online from the United States and India. This cross-cultural comparison is limited in two ways: (1) online cohorts of participants overlap significantly in their exposure to globalized media and (2) comparing only two groups massively underestimates the vast cross-cultural diversity of music and musicality.[56] Despite this, our results revealed significant cross-cultural differences in the evolution of musical structures (see Figures 7, S4, and S5), likely reflecting differences in participants' musical and cultural exposure. Future research could extend our approach to run large-scale cross-cultural experiments that include diverse samples of participants around the world[56] and potentially test the developmental trajectory of emergent song structures by running iterated singing experiments with children.[57]

Transmission mechanisms underlying music evolution

Previous cross-cultural research on human song has focused either on individual psychological processes, investigating music production/perception in highly controlled laboratory settings,[33,58–60] or on large-scale population-level phenomena, analyzing cross-cultural datasets of music recordings or ethnographies.[2,4] These studies have revealed striking differences and similarities in music and musicality cross-culturally, but they have generally overlooked the cultural transmission process. It is thus unclear how population-level structures emerge from individual psychological mechanisms via cultural transmission. By combining these approaches with large-scale iterated singing experiments, we show that social interactions underlying cultural transmission have a substantial explanatory role alongside the contribution of individual biases. In particular, musical structures emerging from our experiments largely depended on whether participants imitated their own or other participants' productions (see Figure 7).

These results can be interpreted as evidence for the violation of at least one of the assumptions of Griffiths and Kalish's model of iterated learning.[19] Namely, either (1) learners do not share the same priors (due to significant individual differences) or (2) there is a failure of the Markov assumption that participants' productions depend only on the data produced by the previous generation (i.e., no learning or context effects). We cannot rule out either of these two explanations.

Regarding the first assumption, our results suggest that participants within each cultural group had consistent melodic priors (see common peaks emerging during individual transmission, Figures 7E and 7F, reflecting shared priors). We also explored the role of individual differences by comparing subgroups of participants based on their self-reported levels of musical experience (see Figure S6), a prominent factor explaining individual differences in singing abilities.[61,62] This analysis revealed some differences but generally similar melodic distributions across subgroups, further suggesting that individual differences alone cannot explain our results.

Regarding the second assumption, it remains a possibility that participants learned their own productions during individual transmissions due to learning or contextual effects. Although our experiments were specifically designed to minimize such effects (see transmission chain designs), further investigation is required to fully understand the mechanisms responsible for differences between social and individual transmission, such as memory and learning, individual differences, and fidelity of copying.

Overall, our results suggest that population-level structures in human song arise from the idiosyncratic biases of individuals that change over time through multiple social interactions. This is consistent with the *interactive* hypothesis,[44,63] which proposes that the structures of human language and music emerge from the underlying dynamics of social interactions that occur during cultural transmission. Over time, this cumulative process generates stable structural compromises that are appealing and easy to learn by all. Here, we support this hypothesis using a highly sophisticated production modality: singing. Both in the United States and India, social transmission rapidly shaped melodies toward homogeneous and simplified structures. This social attractor effect led to larger cross-cultural differences between groups. These results are significant because they provide a new understanding of how social interactions can amplify shared individual biases, contributing to the vast diversity of forms we observe in human song across cultures.[2,4,56] The implications of these results may also be applicable to other behaviors resulting from cultural transmissions, such as language and technology.[14,63]

More broadly, this work demonstrates the benefits of combining large-scale online data collection with innovative psychological paradigms to explore cultural transmission processes in unprecedented detail. The advantage of this approach lies in characterizing the rich collection of oral transmission mechanisms underlying music evolution within a single coherent paradigm that is cross-culturally generalizable. This presents important advances compared with previous iterated learning experiments conducted in the laboratory. For example, it allows us to systematically cover the space of evolutionary possibilities and causally explore the cognitive and environmental factors that influence complex social behavior. Importantly, our study relies on human singing, a key communicative modality for music that is natural and cross-culturally widespread.[2,4,8,62] Thus, our study showcases how large-scale cultural transmission experiments in complex production modalities can transform the kind of research conducted in cognitive science.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Participants
- METHOD DETAILS
  - Automated online implementation

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

- Singing transcription technology
- General procedures
- Experimental paradigms
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Sample-size analysis
  - 1D and 2D Kernel Density Estimation (KDE)
  - Peak finding
  - Trend analysis
  - Melodic contours
  - Comparing distributions
  - Experimental simulations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.02.070.

ACKNOWLEDGMENTS

We thank the members of the Computational Auditory Perception Group, at Max Planck Institute for Empirical Aesthetics, for their help and feedback. We thank Luke Poeppel for contributing to the initial phase of the singing analysis pipeline. We thank Tong Zhao and Elif Çelen for data collection assistance. We thank Ofer Tchernichovski for comments on earlier versions of the manuscript. We thank Richard Widdess and Lara Pearson for useful discussion about Indian melodic structures. For the purposes of Open Access, the author has applied a Creative Commons Attribution License (CC-BY) to any Author Accepted Manuscript arising.

AUTHOR CONTRIBUTIONS

Conceptualization, methodology, software, investigation, analysis, visualization, and writing, M.A.-T., P.M.C.H., and N.J. Project administration, investigation, data curation, M.A.-T. Supervision, N.J. All authors worked collaboratively to discuss methods, analysis, and writing throughout the process of preparing the published work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: November 3, 2022
Revised: December 24, 2022
Accepted: February 23, 2023
Published: March 22, 2023

REFERENCES

1. Brown, S., and Jordania, J. (2013). Universals in the world's musics. Psychol. Music *41*, 229–248.
2. Mehr, S.A., Singh, M., Knox, D., Ketter, D.M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A.A., Hopkins, E.J., et al. (2019). Universality and diversity in human song. Science *366*, eaax0868.
3. Nettl, B. (2010). The Study of Ethnomusicology: Thirty-One Issues and Concepts (University of Illinois Press).
4. Savage, P.E., Brown, S., Sakai, E., and Currie, T.E. (2015). Statistical universals reveal the structures and functions of human music. Proc. Natl. Acad. Sci. USA *112*, 8987–8992.
5. Patel, A.D. (2010). Music, Language, and the Brain (Oxford University Press).
6. Zatorre, R.J., and Baum, S.R. (2012). Musical melody and speech intonation: singing a different tune. PLoS Biol. *10*, e1001372.
7. Hilton, C.B., Moser, C.J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C.M., Simson, J., Knox, D., Glowacki, L., Alemu, E., et al. (2020). Acoustic regularities in infant-directed speech and song across cultures. Nat. Hum. Behav. *6*, 1545–1556.
8. Mehr, S.A., Singh, M., York, H., Glowacki, L., and Krasnow, M.M. (2018). Form and function in human song. Curr. Biol. *28*. 356–368.e5.
9. Merker, B., Morley, I., and Zuidema, W. (2015). Five fundamental constraints on theories of the origins of music. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140095.
10. Savage, P.E. (2019). Cultural evolution of music. Palgrave Commun. *5*, 16.
11. Savage, P.E., Passmore, S., Chiba, G., Currie, T.E., Suzuki, H., and Atkinson, Q.D. (2022). Sequence alignment of folk song melodies reveals cross-cultural regularities of musical evolution. Curr. Biol. *32*. 1395–1402.e8.
12. Tierney, A.T., Russo, F.A., and Patel, A.D. (2011). The motor origins of human and avian song structure. Proc. Natl. Acad. Sci. USA *108*, 15510–15515.
13. Trehub, S.E. (2015). Cross-cultural convergence of musical features. Proc. Natl. Acad. Sci. USA *112*, 8809–8810.
14. Derex, M., and Boyd, R. (2015). The foundations of the human cultural niche. Nat. Commun. *6*, 8398.
15. Centola, D. (2022). The network science of collective intelligence. Trends Cogn. Sci. *26*, 923–941.
16. Tomlinson, G. (2015). A Million Years of Music: the Emergence of Human Modernity (MIT Press).
17. Honing, H. (2018). The Origins of Musicality (MIT Press).
18. Wallin, N.L., Merker, B., and Brown, S. (2001). The Origins of Music (MIT Press).
19. Griffiths, T.L., and Kalish, M.L. (2007). Language evolution by iterated learning with Bayesian agents. Cogn. Sci. *31*, 441–480.
20. Scott-Phillips, T.C., and Kirby, S. (2010). Language evolution in the laboratory. Trends Cogn. Sci. *14*, 411–417.
21. Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: a framework for the emergence of language. Artif. Life *9*, 371–386.
22. Thompson, B., and Griffiths, T.L. (2021). Human biases limit cumulative innovation. Proc. Biol. Sci. *288*, 20202752.
23. Jacoby, N., and McDermott, J.H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. Curr. Biol. *27*, 359–370.
24. Ravignani, A., Delgado, T., and Kirby, S. (2017). Musical evolution in the lab exhibits rhythmic universals. Nat. Hum. Behav. *1*, 0007.
25. Miton, H., Wolf, T., Vesper, C., Knoblich, G., and Sperber, D. (2020). Motor constraints influence cultural evolution of rhythm. Proc. Biol. Sci. *287*, 20202001.
26. Lumaca, M., and Baggio, G. (2017). Cultural transmission and evolution of melodic structures in multi-generational signaling games. Artif. Life *23*, 406–423.
27. Popescu, T., Walther, J., and Rohrmeier, M. (2022). Building blocks of tonality emerge from transmission chains with random melodies. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/vg9fz.
28. Shanahan, D., and Albrecht, J. (2019). Examining the effect of oral transmission on folksongs. Music Percept. *36*, 273–288.
29. Verhoef, T., and Ravignani, A. (2021). Melodic universals emerge or are sustained through cultural evolution. Front. Psychol. *12*, 668300.
30. Lindblom, B., and Sundberg, J. (2014). The human voice in speech and singing. In Springer Handbook of Acoustics (Springer), pp. 703–746.
31. Oxenham, A.J. (2012). Pitch perception. J. Neurosci. *32*, 13335–13338.
32. Pressnitzer, D., Suied, C., and Shamma, S.A. (2011). Auditory scene analysis: the sweet music of ambiguity. Front. Hum. Neurosci. *5*, 158.

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

33. Jacoby, N., Undurraga, E.A., McPherson, M.J., Valdés, J., Ossandón, T., and McDermott, J.H. (2019). Universal and non-universal features of musical pitch perception revealed by singing. Curr. Biol. *29*. 3229–3243.e12.

34. Wong, P.C., Ciocca, V., Chan, A.H., Ha, L.Y., Tan, L.-H., and Peretz, I. (2012). Effects of culture on musical pitch perception. PLoS One *7*, e33424.

35. Krumhansl, C.L., and Kessler, E.J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. Psychol. Rev. *89*, 334–368.

36. Desain, P., and Honing, H. (2003). The formation of rhythmic categories and metric priming. Perception *32*, 341–365.

37. Narmour, E. (1990). The Analysis and Cognition of Basic Melodic Structures: the Implication-Realization Model (University of Chicago Press).

38. Pearce, M.T., and Wiggins, G.A. (2006). Expectation in melody: the influence of context and learning. Music Percept. *23*, 377–405.

39. Dowling, W.J. (1978). Scale and contour: two components of a theory of memory for melodies. Psychol. Rev. *85*, 341–354.

40. Halpern, A.R., and Bartlett, J.C. (2010). Memory for melodies. In Music Perception, M. Riess Jones, R.R. Fay, and A.N. Popper, eds. (Springer), pp. 233–258.

41. Mesoudi, A. (2011). Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences (University of Chicago Press).

42. Burketi, D., and Griffiths, T.L. (2010). Iterated learning of multiple languages from multiple teachers. In The Evolution of Language (Proceedings of the 8th International Conference on the Evolution of Language), pp. 58–65.

43. Navarro, D.J., Perfors, A., Kary, A., Brown, S., and Donkin, C. (2017). When extremists win: on the behavior of iterated learning chains when priors are heterogeneous. Proceedings of the 38th Annual Meeting of the Cognitive Science Society, 847–852.

44. Ravignani, A., Thompson, B., Grossi, T., Delgado, T., and Kirby, S. (2018). Evolving building blocks of rhythm: how human cognition creates music via cultural transmission. Ann. NY Acad. Sci. *1423*, 176–187.

45. Fehér, O., Ljubičić, I., Suzuki, K., Okanoya, K., and Tchernichovski, O. (2017). Statistical learning in songbirds: from self-tutoring to song culture. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160053.

46. Claidière, N., Smith, K., Kirby, S., and Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. Proc. Biol. Sci. *281*, 20141541.

47. Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., and Jacoby, N. (2020). Gibbs sampling with people. Adv. Neural Inf. Process. Syst. *33*, 10659–10671.

48. Marjieh, R., Harrison, P.M., Lee, H., Deligiannaki, F., and Jacoby, N. (2022). Reshaping musical consonance with timbral manipulations and massive online experiments. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.14.496070.

49. Griffiths, T.L., and Kalish, M.L. (2005). A Bayesian view of language evolution by iterated learning. Proceedings of the 27th Annual Meeting of the Cognitive Science Society, 827–832.

50. Langlois, T.A., Jacoby, N., Suchow, J.W., and Griffiths, T.L. (2021). Serial reproduction reveals the geometry of visuospatial representations. Proc. Natl. Acad. Sci. USA *118*. e2012938118.

51. Bowling, D.L., Sundararajan, J., Han, S., and Purves, D. (2012). Expression of emotion in Eastern and Western music mirrors vocalization. PLoS One *7*, e31942.

52. Pearson, L. (2016). Coarticulation and gesture: an analysis of melodic movement in south Indian raga performance. Music Anal. *35*, 280–313.

53. Von Hippel, P. (2000). Redefining pitch proximity: tessitura and mobility as constraints on melodic intervals. Music Percept. *17*, 315–327.

54. Salganik, M.J., Dodds, P.S., and Watts, D.J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. Science *311*, 854–856.

55. MacCallum, R.M., Mauch, M., Burt, A., and Leroi, A.M. (2012). Evolution of music by public choice. Proc. Natl. Acad. Sci. USA *109*, 12081–12086.

56. Jacoby, N., Polak, R., Grahn, J., Cameron, D., Lee, K.M., Godoy, R., Undurraga, E.A., Huanca, T., Thalwitzer, T., Doumbia, N., et al. (2021). Universality and cross-cultural variation in mental representations of music revealed by global comparison of rhythm priors. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/b879v.

57. Trehub, S.E. (2003). The developmental origins of musicality. Nat. Neurosci. *6*, 669–673.

58. Krumhansl, C. (2000). Cross-cultural music cognition: cognitive methodology applied to North Sami yoiks. Cognition *76*, 13–58.

59. Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A.D., and Koelsch, S. (2009). Universal recognition of three basic emotions in music. Curr. Biol. *19*, 573–576.

60. Stevens, C.J. (2012). Music perception and cognition: a review of recent cross-cultural research. Top. Cogn. Sci. *4*, 653–667.

61. Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. PLoS One *9*, e89642.

62. Pfordresher, P.Q. (2022). Singing accuracy across the lifespan. Ann. NY Acad. Sci. *1515*, 120–128.

63. Thompson, B., Kirby, S., and Smith, K. (2016). Culture shapes the evolution of cognition. Proc. Natl. Acad. Sci. USA *113*, 4530–4535.

64. Anglada-Tort, M., Harrison, P.M.C., and Jacoby, N. (2022). REPP: a robust cross-platform solution for online sensorimotor synchronization experiments. Behav. Res. Methods *54*, 2271–2285.

65. Anglada-Tort, M., Harrison, P.M., and Jacoby, N. (2022). Studying the effect of oral transmission on melodic structure using online iterated singing experiments. Proceedings of the 44th Annual Meeting of the Cognitive Science Society, 810–817.

66. Woods, K.J.P., Siegel, M.H., Traer, J., and McDermott, J.H. (2017). Headphone screening to facilitate web-based auditory experiments. Atten. Percept. Psychophys. *79*, 2064–2072.

67. Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam *17*, 97–110.

68. Jadoul, Y., Thompson, B., and De Boer, B. (2018). Introducing parselmouth: a python interface to praat. J. Phon. *71*, 1–15.

69. Duarte, M., and Watanabe, R.N. (2015). Notes on scientific computing for biomechanics and motor control (version V0.0.2). Zenodo. https://zenodo.org/record/4599319#.Y-_LkezP1TY.

70. Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., and Ripley, M.B. (2013). Package 'mass'. R package, version 7, 3.4–51.4.

71. Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423.

72. Drost, H.-G. (2018). Philentropy: information theory and distance quantification with R. J. Open Source Softw. *3*, 765.

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw and analyzed human data	This paper	https://doi.org/10.17605/OSF.IO/UANGD
Model simulations	This paper	https://doi.org/10.17605/OSF.IO/UANGD
Software and algorithms		
R 4.0.5	R Foundation	https://www.r-project.org/
MATLAB R2022a	Mathworks	https://www.mathworks.com/
Python 3.10	Python Foundation	https://www.python.org/
Analysis Code	This paper	https://doi.org/10.17605/OSF.IO/UANGD
Model Simulations	This paper	https://doi.org/10.17605/OSF.IO/UANGD
Singing Technology	This paper	https://doi.org/10.17605/OSF.IO/UANGD
Experimental Code	This paper	https://doi.org/10.17605/OSF.IO/UANGD

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Manuel Anglada-Tort (manuel.anglada-tort@music.ox.ac.uk).

Materials availability

All datasets and analysis code supporting this paper are publicly available as an OSF repository: https://doi.org/10.17605/OSF.IO/UANGD

Data and code availability

All datasets and analysis code supporting this paper are publicly available as an OSF repository: https://doi.org/10.17605/OSF.IO/UANGD

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants

All participants provided informed consent in accordance with the Max Planck Society Ethics Council approved protocol (2021_42). All participants were recruited online using Amazon Mechanical Turk (MTurk). We required four conditions to take part in our experiments: (i) be at least 18 years old, (ii) use headphones or earphones with a working microphone, (ii) be in a quiet environment (e.g., a room with low background noise), and (iii) use an up-to-date Google Chrome browser. These requirements guaranteed good listening conditions and compatibility with our testing platform, PsyNet (see automated online implementation). To ensure high data quality, we only recruited participants with at least 2,000 previously submitted tasks on MTurk with a 95% approval rate on average. Participants were paid at a US $9/hour rate according to how much of the experiment they completed (e.g., if participants failed a pre-screening task and left the experiment early, they were still paid proportionally for their time).

A total of 1,797 participants contributed to the 12 experiments reported in this paper, excluding those who failed pre-screening tasks. Table S1 provides the demographic details of all experiments. Sample size estimation was determined based on a sample-size analysis and is described in quantification and statistical analysis.

METHOD DETAILS

Automated online implementation

PsyNet

All experiments were implemented in PsyNet, a Python package for performing complex online behavioral experiments at large scale.[47,48,64,65] PsyNet is based on the Dallinger framework for hosting and deploying experiments. Participants interact with the experiment via a web browser, which communicates with a back-end Python server cluster responsible for organizing the experiment and communicating with our singing transcription technology. In our experiments, this cluster was managed by Heroku, supporting the experiment management and stimulus generation workload, as well as a Postgres database for sorting results. In those

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

experiments with audio recordings, we also used Amazon Web Services (AWS) S3 storage for hosting stimuli. Code for the implemented experiments can be found in the OSF repository: https://doi.org/10.17605/OSF.IO/UANGD

Recruitment

Participant recruitment was managed by PsyNet. In the iterated learning experiments (Experiments 1–5 and 7–12), participants were recruited automatically until we collected the desired number of chains and a desired length for these chains. The number of chains was determined by a sample-size analysis (see sample-size analysis). In the dense rating experiment (Experiment 6), participants were recruited automatically until we obtained the desired number of ratings per stimulus (the total number of ratings was also determined by a sample-size analysis). Table S2 provides the design parameters of all experiments.

Pre-screening tests

An important source of noise in online experiments is the presence of fraudulent responders, including both computer 'bots' and non-serious respondents, such as participants who do not follow the instructions. We used a combination of techniques to ensure high data quality in online experiments using complex production modalities,[64,65] monitoring performance in real time to provide feedback to participants and exclude fraudulent responders. These techniques were also used to familiarize participants with the main singing procedure. We used three pre-screening tests, presented at the start of each experiment involving audio recordings in the following order: (1) headphone test, (2) audio calibration tests, and (3) singing practice with feedback. For listening experiments, we only used the headphone and volume calibration test.

Headphones test. This test, originally developed by Woods et al.,[66] was intended to ensure that participants were wearing headphones and could perceive subtle sound differences. The test consists of a three-alternative forced-choice task to identify the quietest of three tones. These tones are constructed to elicit a phase cancellation effect, such that when played on loudspeakers the order of quietness changes. Thus, the test can only be passed if participants wear headphones. Participants took six trials and were required to answer at least four trials correctly to pass the test.

Audio calibration tests. The goal of these tests was to ensure that participants were able to record audio in the browser and that the volume level for the input sounds was adequate. The first test consisted of a recording calibration page to check whether we could detect participants' singing input. Since this page required audio recording, Google Chrome automatically triggeres a pop-up window asking participants whether they want to record using their microphone. The experiment could only continue if they accept. Participants were then asked to talk or sing using the microphone to ensure their signal could be recorded, using a sound meter to visually indicate whether the signal had an appropriate sound level (volume). The second test consisted of a volume calibration page. We played a random sequence of tones using our method to generate melodies and asked participants to adjust the volume of the computer to a comfortable level. Participants could pass these tests unless they had technical problems related with audio recording (e.g., having no microphone or not giving permission to record).

Singing practice with feedback. This test was intended to familiarize participants with the singing procedure and ensure they could provide valid signing data. In the first page, we asked participants to sing two tones freely while we recorded them. We played their recording right after, asking whether they could hear themselves (if they could not, we instructed them to make sure they met the technical requirements). On the next page, we provided participants with the main instruction of the singing task:

In each trial, you will hear a melody with 2 notes: Your goal is to sing each note back as accurately as possible. Use the syllable 'TA' to sing each note and leave a silent gap between notes.

After the instructions, participants took two practice trials consisting of hearing a melody and singing it back while we recorded their responses. We used a progress bar to visually indicate the different stages of the recording trial (listening, recording, and finished). The recording was then analyzed using the singing transcription technology (see singing transcription technology). After each trial, we provided feedback indicating the number of tones detected by the technology. In those cases where we could not detect all tones, we provided feedback to participants reminding them of the main technical requirements and instructions of the singing task.

Singing performance test

We developed a singing performance test that served three purposes: (1) measuring participants' general singing abilities, (2) excluding participants who could not provide minimal working data, and (3) detecting participants' singing range (low vs high). The performance test consisted of 10 singing trials, where participants were played two tones and asked to imitate them back by singing, just like in the practice phase. The recordings were analyzed in real-time. Trials were failed if at least one of the three following criteria was met (see failing criteria for details on each performance metric): (1) the number of tones in the target and response were not equal, (2) the maximum absolute interval error ($\text{abs}(I_S - I_R)$) between target interval (I_S) and response interval (I_R) was larger than 3 semitones, or (3) the direction of pitch change in the target interval was different from the direction in the response (in case of unison stimulus we required the response to be between -0.5 and 0.5, as explained in failing criteria). Participants were required to pass at least 5 out of the 10 trials correctly (those participants who did not pass this threshold were excluded from the experiment).

The 10 stimuli of the test consisted of five melodic intervals [-1.3, -2.6, 0, 1.3, 2.6] played at two singing registers (low and high). These interval values were intentionally chosen to avoid any integer semitones which could have primed participants in the subsequent singing task. Moreover, to avoid any effect of tonal context between trials, the tones of the intervals were transposed randomly in each trial according to the center of each singing register condition (see melody generation for details). The order of the trials was randomized for each participant.

To determine the singing register of each participant, we calculated the distance in semitones between the median of the produced tones in the test trials and the center of the low (typical male) and high (typical female) singing registers; in this study set to 49 (or 138.59 Hz) and 61 MIDI notes (or 277.18 Hz), respectively (see melody representation). If the distance between the median produced tone and the center of the low register was smaller, participants were classified in the low-register condition for the subsequent parts of the experiment, and vice versa.

Singing transcription technology

We used a four-step automated process to extract the fundamental frequency (f0) of tones in participants' vocal productions (see Figure 1B). This method is an adaptation of a similar setup extensively used in laboratory experiments with singing.[33] We previously piloted this technology in online singing experiments and found it is highly reliable to extract f0 automatically from audio recordings produced by standard hardware and software available to most participants online.[65] Code for the implemented singing transcription technology can be found in the OSF repository: https://doi.org/10.17605/OSF.IO/UANGD

The first step was to clean the audio signal to remove any artifacts. We empirically found that some recordings contain loud artifacts near the beginning of the recording. We therefore removed the first 100 msec of the recording and added a linear fade-in for another 150 msec. To remove spectral components that do not contain singing information, we then applied a band-pass filter with cutoff frequencies of 80-6000 Hz. Second, we applied an autocorrelation-based pitch estimation algorithm to extract f0 pitch from sung segments,[67] implemented using parselmouth,[68] a Python interface to access Praat. Based on previous piloting with singing data,[65] the autocorrelation method was used with a pitch floor of 65 Hz and pitch ceiling of 622 Hz. The following parselmouth parameters were also modified to non-default values: silence_threshold = 0.03, voicing_threshold = 0.045, octave_cost = 0.03, octave_jump_cost = 0.55, voiced_unvoiced_cost = 0.14.

The third step consisted in segmenting the singing signal into individual isolated tones. We computed the envelope of the amplitude of the singing signal by calculating the maximum square of the amplitude within non-overlapping windows of 20 msec. Then, we used the *detect_peak* function by Marcos Duarte[69] to detect peaks in the envelope corresponding to local maxima of the amplitude (see Figure 1B), which are usually located in the louder sections of a sung tone. Next, we looked for non-overlapping segments of audio that contained multiple peaks. We identified segments that were separated by 70 msec of silence, defined as an audio envelope that is less than -22 db when compared to the first peak identified within a segment (typically at the beginning of the sung tone).

Finally, we filtered the list of segments based on simple heuristics that are useful to remove spurious audio segments (e.g., speaking or background noise). Specifically, we filtered segments with a duration of less than 35 msec. Within segments, we calculated the percentage of duration that exceeded 6 semitones from the median f0 and excluded segments where this threshold exceeded more than 20% of the segment's total duration. The main purpose of this step was to remove octave jumps that occur when the singer's voice "cracks". In addition, a segment with a median pitch outside the allowed range (65-622 Hz) would be excluded. To compute the median f0 for each segment (corresponding to one sung tone), we ignore the first 110 msec and the last 70 msec of the segment, as these often contain less stable singing.

After extracting the f0s from the recordings, we calculated several performance metrics to assess participants' singing accuracy but also to provide feedback in real-time to participants (see automated online implementation) and exclude trials or participants that did not meet basic performance requirements (see failing criteria).

We tested whether the implementation of the singing transcription technology could have introduced some bias to the f0 extraction analysis. We first generated the target audio by sampling pure tones covering the entire pitch space used in our experiments at 0.25 granularity (129 tones ranging in pitch from chromatic 39 to 71 MIDI). We then used the singing transcription technology to extract the f0 frequencies from the audio file. Figure S1 shows the detection accuracy of our technology across the pitch range. The overall mean absolute detection accuracy (difference between target and detected pitches) was 0.0006 (SD = 0.0009; min-max = -0.0009–0.004), demonstrating a high pitch extraction accuracy and only negligible extraction errors.

General procedures

Melody representation

We parametrize the pitch space of melodies as lists of numbers using MIDI notation, which maps each frequency to a positive integer. For example, the middle C in a piano keyboard (C4) is mapped to the MIDI note number 60. Formally, the frequency-to-MIDI mapping is given by:

$$f = 440 \cdot 2^{(m-69)/12} \quad \text{(Equation 1)}$$

where m is the MIDI note and f is a frequency measured in Hz. MIDI notation is useful to express absolute pitches in a logarithmic scale, which is shown in previous work to represent the nature of pitch perception.[33] Thus, using MIDI numbers we can represent melodies in *absolute pitch representation* (e.g., [64.12, 58.79, 63.24]) while also studying frequencies with high granularity, including intermediate fractal values beyond the discrete frequencies typically used in the Western 12-tone equal temperament scale (the keys in a regular piano keyboard). However, most people represent melodies using *relative pitch representation*,[39] where pitches are expressed relative to each other rather than in absolute terms. Thus, melodies can also be represented as a sequence of intervals, expressing each pitch relative to the previous one (in semitones; e.g., [-5.33, 4.45]).

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

Melody generation

Previous research on melody cognition has traditionally used discrete musical systems, mostly consisting in the Western 12-tone chromatic scale.[35,39,40,58] Consequently, results obtained from this research are inherently biased towards Western music rules, neglecting the use of microtuning (e.g., intervals smaller than a semitone) and other tuning systems beyond equal temperament which is fundamental to many non-Western music traditions, such as the use of quartertones and 1/8th tones in Arabic and Turkish music. A key feature of our method is that it does not assume any culturally specific knowledge about discrete scale systems *a priori*. Instead, we take advantage of novel psychological testing methods (see automated online implementation) to study vast continuous melodic spaces with high resolution - i.e., sampling melodic intervals or pitches continuously from a uniform distribution (see Figure 1D). This method allows us to ensure that any effects observed in transmission chain experiments can be attributed to human reproduction biases rather than to constraints imposed by the stimulus generation process. This method is also applicable to individuals from any musical or cultural background.

Uniform pitch sampling. In most of our experiments (Experiment 1-3, 6-12, we used a *uniform pitch sampling method* to generate melodies. This sampling method consisted of three steps. First, we obtained a singing register for each participant. In the singing experiments (Experiment 1-3, 5, 7-12), this was determined automatically based on the participants' performances in the singing test completed at the start of the experiment (see singing performance test). Based on common singing registers used in previous work,[33] we set the center of the low and high registers (c) to 49 and 61 MIDI notes, respectively, thereby separating the centers of the two registers by an octave. In the listening experiments (Experiment 4 and 6), we set a middle register for all participants with a center at 55 MIDI. Second, we used a roving technique (see pitch roving procedure) to obtain a virtual reference pitch for each melody by uniformly sampling a real number around the center of the singing register (r). Finally, we obtained each tone in the melody by uniformly sampling a real number with a fixed pitch range centered on the reference tone (thus the reference tone was not directly played). The pitch range consequently determined the max range of the pitch space in the initial set of melodies in each experiment (see Table S2 for the exact parameters used in all experiments). However, participants were allowed to reproduce melodies outside this range (see failing criteria). Formally, melody tones (t_i) were randomized with the following formula: $t_i = c + r + n_i$, where c is the center of the singing register, r is the roving value (sampled uniformly), and n_i is the relative pitch value sampled uniformly within a fixed pitch range.

Randomizing pitches uniformly is particularly useful when generating longer melodies because it guarantees that the distribution of each tone (T_n) is uniform and identical regardless of melody length, and it also guarantees that the distribution of melodic intervals (the difference between consecutive tones) is identical regardless of melody length. For example, the first interval (the difference between the second and first tones in the melody) will have the same distribution as the fourth interval (the difference between the fifth and the fourth tones). Formally: $T_2 - T_1 \sim T_5 - T_4$, where T_n is the random variable associated with the n^{th} tone. However, the distribution of consecutive melodic intervals using this method is not uniform. Instead, consecutive intervals obtained from uniform pitch sampling will have the distribution of the convolution of two uniform variables (higher probability density in the center of the distribution and linear decays away from the center). Thus, combinations of large consecutive intervals (e.g., [18, 18] semitones) are unlikely. It is not possible to generate long melodies, so they have both uniform pitch and interval distributions, so we chose uniform pitch sampling because it has more advantages – i.e., identical (uniform) distribution of pitches and identical (nonuniform) distribution of consecutive intervals.

Uniform Interval Sampling. When studying single melodic intervals (two tones played sequentially), there is no need to worry about the distribution of consecutive intervals. Thus, for those experiments using iterated learning with one-interval only (Experiment 4-5), we used instead a *uniform interval sampling method* to generate stimuli. This procedure also consisted of three steps. First, we sampled intervals uniformly within a range of [-15, 15] semitones. Second, we used a roving technique (see pitch roving procedure) to obtain a reference tone for each melody by uniformly sampling a real number around the center of participants' singing register (r). Finally, we obtained the two melody tones by using the reference tone as the first tone and the reference tone plus the interval as the second tone. Formally, the two tones here were given by: $t_1 = c + r$ and $t_2 = c + r + I$, where c is the center of the singing register, r is the roving value, and I is the interval (sampled uniformly within a range of [-15, 15] semitones).

Pitch roving procedure

When studying melody perception and production, there are important subtleties in the generation of melodies that should be addressed. First, there is the risk of generating contextual effects of implied tonality across melodies, where a given trial may be interpreted based on the tonal context from the preceding trial. Second, we found that it is important to consider participants' singing range. For example, asking participants to reproduce melodies outside their singing range would significantly increase the complexity of the task, involving different cognitive mechanisms such as octave equivalence,[33] and likely introducing additional production noise. Indeed, it has been shown that non-musicians may struggle with octave equivalence (the transposition of melodies to a comfortable singing range), and individuals differ largely in their ability to transpose melodies relying on octave equivalence.[33] In our experiments, it was important to avoid such situations for both practical reasons (e.g., avoiding extreme responses and audio artifacts) and ethical considerations (e.g., requiring participants to use their voice uncomfortably for prolonged time).

To address this, we used a roving technique to randomize the absolute pitches of melodeis in each singing trial, thereby only transmitting intervallic information across generations (relative pitch representation). This procedure consisted of two steps. First, we computed the difference between each absolute pitch in the target melody and the reference tone used to generate them: $n_i = t_i - (c + r)$. Second, we obtained a new reference tone by uniformly sampling a real number within a roving width of ±2.5 semitones around the center of participants' singing register (the singing register was determined automatically in a singing

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

performance test prior to the main task, see singing performance test). We used a roving value of ±2.5 semitones because it was found in previous singing experiments to provide a good tradeoff between values that were (1) not too small (so tonal context between trials can be effectively removed) or (2) not too large (so that vocal range is not altered significantly).[33] Third, we used the randomized reference tone (r') to create a new sequence of tones (t'_i) is: $t'_i = c + r' + n_i$, either using uniform pitch or interval sampling (see melody generation).

This roving procedure preserves intervallic information: $t'_i - t'_j = (c + r' + n_i) - (c + r' + n_j) = t_i - t_j$, while resampling the absolute pitch information. This means that successive melodies will typically correspond to different musical scales, and participants will therefore be encouraged to hear each melody on its own terms instead of relating it to the previous melody. In other words, it minimizes the implied tonal context between consecutive melodies. This technique also ensures that all melody tones remain within a comfortable singing range. For example, if the max pitch range parameter used to sample tones is set to 10 semitones (like in Experiment 1), the maximum possible pitch range in all melody tones (the intervals between the reference pitch and the highest or lowest possible pitch) will be 20 semitones (see Table S2 for the pitch range parameters used in all experiments). By using the roving technique described above, we can make sure that this maximum pitch range of 20 semitones is centered around participants' singing register throughout the experiment. Moreover, this technique allows us to combine participants with different singing ranges within the same transmission chain.

Melody presentation

Melody tones in all experiments were 550 msec long and separated by 250 msec of silence (inter-onset-intervals were 800 msec long). Tones were synthesized using complex tones containing 10 harmonics with 14 dB per octave exponential roll-off. Each tone started with a 20 msec exponential raise in amplitude (attack), followed by a 50 msec decay to an amplitude of 0.8 of the maximum amplitude. This followed by an exponential release of additional 480 msec. Thus, the overall duration of the sound was 550 msec. We played the stimuli using Tone.js, a Web Audio framework for generating sound in the browser.

Failing criteria

We used different failing criteria depending on the requirements of each experimental task to monitor participants' performance in real time, provide feedback, and exclude participants who did not provide valid responses. In the singing trials, the failing criteria was based on the output of participants' recordings calculated after each trial (see singing transcription technology). Table S2 specifies the exact failing criteria used in each experiment.

Correct number of tones. For all singing experiments, the number of detected tones in each response was calculated and compared against the number of tones in the target melody. As a universal failing criterion, trials failed when the number of detected tones in the response did not match the number of tones in the target melody. The only exception was Experiment 3, where we relaxed this failing criteria to allow participants to make small errors in the number of reproduced tones depending on the length of each melody (if melodies had less than 5 tones, we allowed a difference of 3 tones between target and response, if melodies had 6-8 tones, we allowed a difference of 4 tones, if melodies had 9 or more tones, we allowed for a different of 5 tones).

Max absolute interval error. In some experiments (Experiment 1, 10-12), it was fundamental to ensure that singing productions were accurate, so any differences in the results could not be explained away by differences in singing performance among participants. For example, this was particularly important when comparing different modes of transmission (across vs within) or participants from different cultural backgrounds. Thus, in these experiments we used an additional performance metric to fail any trials that were not accurate in terms of the distance between the target melodic intervals and sung response. In particular, we defined a max absolute interval error as, $E=|I_s-I_r|$, where I_s and I_r are the stimulus and response intervals, respectively. We then failed any trial where the max absolute interval error was larger than 5.5 semitones, effectively ensuring some degree of acurracy when copying melodies. We found that 5.5 semitones was a good arbitrary threshold to ensure accuracy while not losing too many trials.

Max pitch range. In those singing experiments using uniform pitch sampling (Experiment 1-3, and 6-12), we also excluded trials in which detected pitches were two times larger than the max pitch range used to sample melody tones. Note that max pitch range determines the maximum possible interval between the reference tone used to generate melodies (randomized and centered around participants' singing range) and each tone in the melody (see melody generation and pitch roving procedure for details). Thus, if the max pitch range was set to 10 semitones, we only excluded trials with pitches that were higher or lower than 20 semitones relative to the reference tone of the melody (twice the size of max pitch range). Consequently, this would allow any melodic intervals within a range of [-40, 40] semitones. This criteria is very relaxed, as it allowed responses well above the singing capacities of non-professional singers, ensuring that participants could freely use the pitch and intervallic space, but avoiding the detection of sound artifacts from background noise at very low and high frequencies (see Table S2 for max pitch range parameters used in all experiments).

Max interval size. In those experiments using uniform interval sampling (Experiment 4-5; see melody generation), we excluded trials that were outside the interval range used to generate melodies - i.e., participants' responses that were larger or smaller than [-15, 15] semitones. This step guaranteed that all productions remained within the intervallic space under study. Note, however, that this failing criterion was only used in the two experiments using uniform interval sampling (Experiment 4-5), as these were the only experiments transmitting melodies composed of one interval (two pitches).

Direction accuracy. In the singing performance test (see singing performance test) and singing feedback trials (see pre-screening tests), we also measured whether the direction of pitch change in the target interval of the test was the same as the direction in participants' responses. The direction of pitch change in each interval was categorized either as ascending (the second tone is larger than 0.5 semitones compared to the previous tone), descending (the second tone is smaller than 0.5 semitones compared to the previous tone), or the same (the first and second tone were both within 0.5 semitones distance).

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

In experiments using across-participant chains, when a trial failed, a new participant was allocated to that trial until a valid response was given. In experiments using within-participants chains, the same participants were allocated to the failed trial until a valid response was given. To avoid fatigue in the within-participants experiments, we only allowed four failed trials per chain in each generation. If there was an instance exceeding this threshold, the entire chain was excluded from the experiment.

Validation of general procedures

We tested whether the general procedures described above – melody generation, pitch roving procedure, failing criteria – could have introduced any significant biases to the behavioral results reported in the paper by running a simulation study. Using the same code employed in the transmission experiments, we sampled 1,000 random melodies (composed of 3 tones) and simulated the iterated learning procedure by transmitting these melodies across 100 generations. We used Gaussian noise (M = 0, SD = 1) to copy the input melodies in each generation. The results of the simulation are shown in Figures S1B and S1C, including both the joint marginals of the melodies and melodic features across generations. The results demonstrate that there are no biases in the implementation of the general procedures that could account for the behavioral results obtained in our experiments. Indeed, the simulated data shows trends in the opposite direction of what we observed when running the experiments with human participants (e.g., a gradual increase in interval entropy and mean abs interval size over generations). This occurs because adding gaussian noise in each generation gradually increases the distribution to be more uniform, thereby increasing interval entropy and mean absolute interval size. The results of this simulation are interesting because they show how the data derived from our experiments could look in the absence of any systematic reproduction errors in our participants.

Experimental paradigms

We used a combination of experimental paradigms, consisting of variations of iterated learning and listening experiments. The main design parameters for each experiment are summarized in Table S2.

Iterated singing

Experiments 1-3, 5, and 7-12. To study melodic transmission at a large scale, we developed an automatic online pipeline that streamlines transmission chain experiments in the singing modality. Participants are initially presented with a random "seed" melody (a sequence of tones randomly generated from a continuous space) and asked to reproduce it by singing (Figure 1A). Their reproductions are then synthesized on the fly to generate the stimuli for the next participants (Figure 1B; see singing transcription technology). Over generations, participants' production biases are amplified, allowing us to measure human reproduction biases.

Iterated singing experiments were initialized by selecting a desired number of chains per experiment (see quantification and statistical analysis). Chains were then initialized by randomly and continuously sampling each melody tone from a uniform distribution (see melody generation). Chains evolved for 10 generations, as we found this length to provide a good tradeoff between feasibility of data collection and convergence of the results. Our experiments covered most of the singing range, with a maximum interval range of [-20, 20] semitones in experiments with three-tone melodies (Experiment 1, 7-12) and a slightly reduced range of [-15, 15] semitones in experiments with more complex transmission tasks (Experiment 2-5), such as transmitting longer melodies or using sliders. Participant and experiment details are summarized in Tables S1 and S2.

From the participants' point of view, all singing experiments were the same. All singing trials consisted of hearing a synthesized target melody and singing it back as accurately as possible (see procedure for further details). Participants were not aware that they were interacting with other participants or that they were taking part in a transmission chain experiment.

Transmission chain designs. A design feature in our experiments concern the way in which participants acquire information from the previous generation. In particular, melodies can be either transmitted using *across-participant chains* or *within-participant chains* (Figure 1C). In across-participant chains (social transmission), participants copy melodies produced by other participants. In within-participant chains (individual transmission), the entire chain is completed just by one participant. We used across-participant chains in most of our experiments (Experiment 1-5, 7-10) because our goal was to study cultural transmission, where melodies are passed from one particiapnt to the next. However, we implemented within-participant chains in two experiments (Experiments 11-12) to study the effect of social versus individual transmission.

Across-participant chains (Experiment 1-5, 7-10). Collecting data in across-participant chains is practically harder because completing an entire chain depends on the interactions between multiple participants (e.g., if all participants provide 10 trials, an experiment would need 100 active participants within one experimental session in order to complete 100 chains with 10 generations). Thus, the pool of active participants within a single experimental session determines how many chains can be completed. When running singing experiments with short melodies and recruiting online participants from the US (Experiment 1), we were able to complete about 200 chains with 10 generations (a total of 2,000 singing trials) within a single experimental session, requiring approximately 50-60 participants per session and 40 trials per participant. However, in more complex singing experiments, such as with longer melodies (Experiment 2 and 3) or using a memory perturbation task (Experiment 7-9), we required 30 trials per participant because the trials were longer, consequently reducing the number of completed chains per experimental session. Thus, to obtain the desired number of chains per experiment, we often ran multiple experimental sessions (e.g., in Experiment 1, we deployed three experimental sessions to obtain a dataset of about 600 chains). In such cases, the same participants could participate more than once in different experimental sessions. However, participants were not allowed to contribute to more than 3 experimental sessions in any of the experiments.

Within-participant chains (Experiment 11-12). Collecting data in within-participant chains is comparatively easier because each chain only depends on the performance of one participant. However, transmitting melodies in within-participant chains has a

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

potential confound of memory that does affect across-participant chains: since participants take all trials in each chain, it is possible that they learn their own previous productions.

To minimize potential memory confounds, participants always completed four transmission chains in parallel (e.g., we required 40 singing trials per participant, so participants could complete 4 full chains with 10 generations each). This allowed us to intersperse the singing trials across the four chains and make sure that trials from the same chain could not be taken in succession (e.g., participants would need to take at least three trials from different chains before repeating a trial from the same chain). Second, the roving procedure described above (see pitch roving procedure) was intended to remove any tonal context between trials, thereby removing any possible memory effects based on absolute pitch representations.

Procedure. After completing the pre-screening tests and singing performance test (see automated online implementation), participants were presented with the instructions of the singing task:

In each trial, you will hear a melody with [NUMBER OF NOTES] notes. Your goal is to sing each note back as accurately as possible. Use the syllable 'TA' to sing each note and leave a silent gap between notes.

They were also informed that their performance would be analyzed after each trial. Participants then took three practice trials and were given verbal feedback after eacah trial based on their performance (using the "failing criteria" described above). After the practice phase, participants completed a singing test consisting of 3 further trials. This time, we excluded participants based on their performance. Participants could only pass the test if they completed at least 2 trials correctly. This filtering step was intended to ensure that data quality in the main singing task was high. In the main singing task, participants were reminded of the main instructions one more time, and then started with the main experimental block consisting of 30 or 40 singing trials, depending on the experiment. In each trial, participants were randomly allocated to one of the parallel transmission chains available in the experiment. Moreover, to help participants record their singing productions, we implemented a progress bar to visually indicate the different stages in the recording: listening and singing. At the end of the experiment, participants filled the generic demographic questionnaire (see questionnaire).

Cross-cultural comparison. We compared the results of iterated singing experiments between US participants (Experiment 1 and 11) and Indian participants (Experiment 10 and 12). Participants' nationality was selected using the MTurk qualification system. Participants in the two experiments had similar demographic compositions and levels of musical expertise (see Table S1).

Online recruitment of Indian participants was practically harder because the pool of active MTurk participants in this location is significantly smaller. In within-participant chains (Experiment 12), this was less problematic because we could deploy multiple experimental sessions in different days. However, we still needed a total of 7 experimental sessions to recruit about 73 participants (collecting data from 10 participants approximately in each session). To be consistent with the other experiments, we did not allow participants to contribute to more than 3 experimental sessions. However, online recruitment is more constrained in across-participant chain experiments, as it requires data from multiple participants (see transmission chain designs). To collect the dataset for the across-participant chain experiment in India (Experiment 10), we ran the same experiment on different days, collecting data from new participants until the desired number of chains was completed. In this method of deployment, we collected as much data as possible per session and when there were no more active participants left, we exported the data. We then deployed the experiment on a different day to recruit new participants but started with the data we previously obtained. Note that according to the sample-size analysis we initially aimed to recruit 200 chains, but due to the these recruitment constraints, we were able to collect a total of 120 completed chains and 54 unique participants by running the experiment incrementally in three different sessions. In this experiment, participants were not allowed to participate in different sessions.

Iterated slider imitation

Experiment 4. This experiment measured the effects of transmission using an iterated learning paradigm where participants copied melodies with a slider (instead of their voice). We made a number of changes in the iterated paradigm to make this task feasible with a slider. First, we decreased the complexity of the stimulus space by only studying one-interval melodies. The rationale behind this decision was to represent all melodies in a one-dimensional horizontal line, where each location represents a unique interval in the space. In this way, participants could match the target interval simply by moving a slider horizontally along the plane, keeping the first tone constant within each trial (see procedure). Second, we used an aggregation technique to reduce noise in experiments using sliders while maximizing subtle perceptual effects.[47] Namely, we aggregated the slider responses of three participants in each trial and passed the median answer to the next generation. The result was that the aggregated copying error for the slider was closer to the unaggregated copying error for the singing control experiment (Experiment 5). Figure 4C shows the trends in copying error in the singing (unaggregated) and slider (aggregated) experiments. The results are more comparable due to the aggregation technique, even though imitating melodies with sliders was still harder.

Procedure. After completing the pre-screening tests (headphones and volume calibration tests), participants were presented with the main instructions of the task:

At the beginning of each trial, you will be played a particular melody. Your task will then be to copy that melody with your slider. You will be awarded performance bonuses depending on how well you answer each question. You will only hear the reference melody once, so listen carefully to do as well as possible!

Participants then took three practice trials, receiving feedback on their performance along with their bonus. Participants then started the main task consisting of 60 trials. In each trial, participants were randomly allocated to one of the parallel transmission chains available in the experiment and asked to imitate the target melody by moving a slider. Releasing the slider triggered the new interval to be played corresponding to the updated position. To ensure participants explore the horizontal space minimally,

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

we required that they interacted with the slider at least three times. Note, however, that the target melody only was played once at the start of the trial. When participants were satisfied with their response, they clicked the "next" button to move to the next trial, storing the slider location as the answer. Participants received feedback after each trial with a performance incentive depending on how accurate their response was (see performance incentive). At the end of the experiment, participants were provided with the generic demographic questionnaire (see questionnaire).

Performance Incentive. Since copying melodies with a slider is comparatively less intuitive than singing and there is no research reporting copying accuracy in this modality, we avoided using any failing criteria in this experiment. To incentive participants to perform the task honestly, we used a performance incentive strategy rewarding participants who performed accurately. In each trial, we calculated the distance between the target melody and the given response using the mean absolute interval size: abs($I_S - I_R$). We then assigned a financial bonus (in dollars) to each response by:

$$B = B_0 \cdot max(0, 1 - |I_S - I_R| / T) \quad \text{(Equation 2)}$$

where B_0 is 0.05 US dollars and the threshold $T = 1$ semitones. We ensured the result was always positive, and if the participant error was larger than the threshold, no bonus was given. After each trial, the resulting bonus was presented to participants in a feedback page also providing verbal feedback based on the performance: "excellent" (if the absolute interval error was smaller than 0.25 semitones), "good" (smaller than 1 semitone), or "bad" (larger than 1 semitone). We found this procedure to work well in subjective rating experiments performed online.[47]

Dense rating paradigm

The dense rating paradigm has been previously used to characterize the subjective pleasantness of harmonic intervals (tones played simultaneously) covering vast stimuli spaces with high resolution.[48] Here, we applied this paradigm for the first time to evaluate the aesthetic appeal of melodic intervals (tones played sequentially) densely sampled from a continuous intervallic space of range [-15, 15] semitones. In this paradigm, participants heard intervals that were randomly and densely sampled from a continuous intervallic space (e.g., 1.87, 12.33, or 4.52 semitones), using the same procedure to generate melodies (see melody generation). Each sample received a pleasantness rating on a scale from 1 (completely disagree) to 7 (completely agree). Based on a sample-size analysis, we found that 15,000 samples were sufficient to reliably estimate the distribution of melodic pleasantness (see "sample-size analysis"quantification and statistical analysis).

Procedure. After completing the pre-screening tests (headphones and volume calibration tests), participants were presented with the main instructions of the task:

In each trial, you will be presented with a word and a sound. Your task will be to judge how well the sound matches the following word: ***pleasant***.

Participants then started the main task consisting of 60 trials. In each trial, participants were played a sound and instructed to rate it using the 7 choice options. At the end of the experiment, they filled out the generic demographic questionnaire.

Questionnaire

At the end of each experiment, participants were provided a general questionnaire asking for basic demographic information (i.e., nationality, self-reported gender identity, years of musical training, and overall feedback about the experiment). We also collected general information about participants' musical expertise using the standardized Musical Training factor from the Gold-MSI test.[61]

QUANTIFICATION AND STATISTICAL ANALYSIS

Sample-size analysis

We performed a sample-size analysis to estimate the necessary number of transmission chains in iterated singing experiments to produce stable estimates at two levels of analysis: (1) the distribution of singing responses in the last three generations of the experiments and (2) changes in structural melodic features aggregated across generations. First, we used the data of Experiment 1 to assess the stability of the results in these two levels as a function of the number of chains per dataset, ranging from 50 to 600 chains in bins of 50 chains. In each bin, we estimated the stability of the distribution of responses by randomly splitting the dataset by half and calculating the JSD of the two split distributions (see comparing distributions for details), a measure of overall similarity between probability distributions. Second, to estimate the stability of the trend analysis in melodic features aggregated over generations in each bin, we calculated the Pearson correlation between the aggregated interval entropy (see interval entropy) over generations in the two split datasets using the *cor.test* function in *R*.

Figure S7 shows the results of the sample-size analysis, indicating that 600 chains were sufficient to reliably estimate the distribution of responses (*JSD* = 0.01), whereas 150-200 chains were sufficient to reliably estimate temporal changes in structural melodic features aggregated across generations ($r = 0.8$). Thus, we aimed to collect 600 chains for the two main singing experiments with short melodies in the US (Experiment 1: 590 chains; Experiment 11: 615 chains). For the more complex iterated learning experiments, collecting high-powered datasets would have been practically harder and costly, so we aimed to collect 150-200 chains. This includes the iterated singing experiments with longer melodies (Experiment 2-3), the iterated learning experiments using a slider and a control comparison (Experiment 4-5), and the memory interference experiments (Experiment 7-9). The number of chains in the experiments with Indian participants (Experiment 10: 120 chains; Experiment 12: 223 chains) was determined based on the maximum possible number of active participants we were able to recruit online (see cross-cultural comparison).

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

We ran a second sample-size analysis for the dense rating experiment (Experiment 6), using the same procedure described above to measure the reliability of the distribution of responses, but this time as a function of the number of stimuli sampled from the space. The results indicated that 15,000 stimuli were sufficient to reliably estimate the distribution of responses (Figure S7; *JSD* = 0.0003).

1D and 2D Kernel Density Estimation (KDE)

We used a Gaussian kernel smoother to summarize the results from all trials in the iterated learning experiments and the dense rating paradigm. The 1D and 2D KDEs were computed over a grid of 1,000 points spanning the interval range of interest (see Table S2 for experiment details). For all statistical analysis and 1D KDEs, we used a bandwidth of 0.25 semitones (a quarter tone). We found this parameter to provide an adequate degree of smoothing to maximize interpretability. To estimate uncertainty, we computed the bootstrapped standard error for the smoothed values through nonparametric bootstrapping with 1,000 replicates. In the iterated learning experiments, we bootstrapped with replacement over transmission chains, whereas in the dense rating experiment (Experiment 6) we bootstrapped over participants. For the 2D KDEs visualizations (Figures 2A and 2B, 5B, 6A and 6B, 7A and 7B, S3, and S4), we used the default bandwidth estimator implemented in the function *kde2* from the R package MASS[70] but adjusted by a factor of 0.5 for greater precision. For visualization purposes, we plot all 2D KDEs using a range of [-12, 12] semitones. To aid interpretability across datasets, 2D KDEs are expressed relative to a uniform distribution, where dark blue indicates low density and yellow indicates high density.

Peak finding

To identify statistically reliable peaks from behavioral data, we use the following steps: (1) we created 1,000 bootstrapped datasets by sampling the last three generations of the input data with replacement over the chains; (2) for each bootstrap dataset, we computed the kernel density estimate associated with the data (we used a bandwidth of 0.25 semitones in all analysis); (3) we then computed all peak locations for a given bootstrapped dataset using MATLAB's *findpeaks* function with default parameters. To identify the average number of detected peaks in each generation (see interval vocabulary size), we used the same procedure but identified peaks in each generation of the input data separately and averaged the resulting number across bootstrapped datasets.

In some analyses, we were interested in estimating the reliability of these peaks in order to identify meaningful peak clusters or interval categories (marked as red dots and shaded areas in Figures 2, 4, 6, and 7). This procedure had five steps:

1. We took the 1,000 bootstrap replicates of the kernel-smoothed behavioral profiles created previously, and ran the peak finding algorithm on each of these, producing 1,000 sets of peak estimates.
2. To identify interval categories in each set of peak estimates, we computed the kernel density estimate of the peak locations (with the same bandwidth of 0.25), rather than using the data directly. The resulting peak-density distribution corresponds to the probability of finding a peak near a given interval. We then identify the peak categories by using Matlab's *findpeaks* on the average peak-density distributions over the bootstrapped datasets. This non-parametric method allowed us to estimate peak category locations while avoiding the use of predefined sets of interval categories.
3. For each bootstrapping dataset and interval category, we then found the closest peak within a +/- 0.5 semitones window (if such a peak existed).
4. We counted the proportion of bootstrap iterations where a peak was observed within that window. We considered that a peak was statistically reliable if the proportion was greater than 90%. We found this threshold to achieve a good balance between reliability and interpretability across all datasets.
5. Finally, we calculated the mean location of the bootstrapped peak locations and 95% CI by averaging the peaks associated with that location.

Trend analysis

In all trend analyses conducted in this paper (see Table S3), we used linear regressions with 95% confidence intervals derived from bootstrapping over chains (1,000 replicates, Gaussian approximation). For each bootstrapped dataset, we obtained the coefficient *B* indicating the slope of the regression line (the regression model was implemented using the *lm* function in R). We then averaged the *B* coefficient of all bootstrapped datasets and obtained the final coefficient B reported in Table S3 along a measure of uncertainty using 95 % CI. Thus, the resulting *B* represents the change in the dependent variable associated with each trend analysis and if the CI does not include 0 (indicating no linear relationship), we infer that the relationship is statistically significant (the sign of *B* indicates the direction of the trend). We measured four core melodic features: interval entropy, interval vocabulary size, interval size, and copying error.

Interval entropy

To quantify the complexity of melodies using an appropriate summary statistic, we calculated the entropy (*H*) of the distributions of intervals across all melodies in each generation $I_t \sim p(I_t)$ using Shannon's formula[71]:

$$H(I_t) = -\int p(i_t) log_2 p(i_t) di_t \quad \text{(Equation 3)}$$

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

To compute interval entropy, we discretized all intervals at 0.25 semitone granularity and computed this integral numerically with a discrete grid.

Interval vocabulary size

We measured the interval vocabulary size by computing the average number of peaks in the distribution of intervals in each generation. We identified peaks using the procedure described above (see peak finding).

Interval size

To quantify the average interval size of melodies across generations, we calculated the absolute mean interval size (a) of each melody:

$$a = \frac{1}{N}\sum_{n=1,\ldots,N}|I_n| \quad \text{(Equation 4)}$$

where I_n are the N intervals of the melody. We then calculated the mean absolute interval size of all melodies in each generation.

Copying error

We computed copying error as the root mean square distance between the target melody and response: Formally, let I_{S_n} and I_{R_n} be the stimulus and response intervals of the n^{th} tone of a melody of length N. We define the copying error (e) of a melody as follows:

$$e = \left(\frac{1}{N}\right)\sqrt{\sum_{n=1,\ldots,N}\left(I_{R_n} - I_{S_n}\right)^2} \quad \text{(Equation 5)}$$

We then calculated the average of e overall all melodies in each generation.

Comparing melodic features between experiments

When comparing the evolution of melodic features between different experiments (Figures 3A, 4C, 5C5F, 6C–6F, and 7C, 7D), we linearly normalized the features based on the baseline values at the start of the experiments. This is because the experiments differed in key design parameters as well as the total number of transmission chains (see Table S2), which caused them to differ in their starting baseline levels in different melodic features. For example, the mean absolute interval size in Experiment 1 (three-note melodies) was higher than the mean interval size in Experiment 4 (five-note melodies) because the maximum pitch range used to sample melodies was larger in the former (20 semitones) than in the later (15 semitones). Such differences also generated different starting values in the entropy of the interval distribution and the number of detected peaks. Thus, we used this strategy to directly compare changes in melodic features' trends between experiments. The only exception was copying error, where we kept the absolute values across all figures. This is because copying error has no values at the start of the experiment (generation 0) and because it is important to compare error in absolute terms, so we can interpret the overall difficulty across experiments.

Melodic contours

We used two steps to calculate the average melodic contours of melodies in Experiment 2 and 3 (Figures 3B, 3C, and 3F). First, we aligned the range of all melodies, by transposing each melody to the high singing register so that melodies in the low register were transposed an octave above. Second, for each tone in the melody, we calculated the mean MIDI value (and SE) across all melodies in each generation.

For the clustering analysis (Figure 3C; Experiment 2), we transposed the melodies to the high singing register and also centered them by subtracting the average pitch of each melody from each melody tone. We then applied k-means clustering to the five notes of the melodies in the last three generations of the experiment. To visualize the clusters, we performed a PCA on all note melodies and used the two main components (explained variances of 56% and 29%, respectively) to plot all melodies along the two-dimensional space, where dots represent melodies and the color represents their cluster.

Comparing distributions

For two distributions P, Q we compute the Jensen-Shannon divergence (JSD) as follows, implemented using the *JSD* function from the *philentropy* package in *R*[72]:

$$JSD(P,Q) = \frac{1}{2}D(P,M) + \frac{1}{2}D(Q,M) \quad \text{(Equation 6)}$$

Where $M = \frac{1}{2}(P+Q)$, and

$$D(P,Q) = \int p(x)log_2(p(x)/q(x))dx \quad \text{(Equation 7)}$$

Note that the JSD is symmetric and always between 0 and 1, and that the JSD of two distributions is 0 when the two distributions are identical. The JSD obtains the maximal value of 1 when the two distributions have non-overlapping support (regions with probability larger than 0). In practice we compute the JSD numerically over a discrete grid.

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

Experimental simulations

Interval-size model

To simulate simple constraints that only depend on interval size and direction, we used a model where the response interval I_R depends only on the previous interval size and direction (I_S) using a simple polynomial function b, and an additional independent gaussian noise n. Formally:

$$I_R = I_S + b(I_S) + n \quad \text{(Equation 8)}$$

To find the function b, we used data (I_S and I_R) from the first generation of Experiment 5 (iterated singing with two-tone melodies) and computed b', a 7th-order polynomial function fitted with Matlab's *polyfit* with default parameters (see Figure S2 for the resulting polynomial function derived from empirical data). Since the overall magnitude of the bias changed slightly across generations, we used a single scalar parameter to scale the fitted polynomial function (the same parameter and bias function was used across all generations), namely $b(I) = c \cdot b'(I)$. The free parameters of the model are thus the noise magnitude $\sigma' = std(n)$ and the bias scaling c. The parameters are optimized as described below (see model performance and parameter optimization). Figure 4G shows the aggregated results of the model in the last 3 generations of the simulation (Figure S2 shows an example of the model across generations). The code for the simulations is part of the OSF repository associated with this paper: https://doi.org/10.17605/OSF.IO/UANGD

Preference model

To simulate the role of subjective preferences on melodic transmission, we used a model for serial reproduction that translates preferences (a subjective utility function) to a perceptual prior (see the development of the model in the appendix of Harrison et al.[47]). The model takes as input the data from the subjective preference experiment (Experiment 6; see Figure S2 for the aggregated function derived from empirical data) and predicts all data from the following iterations based on a standard serial reproduction Bayesian model with this function as a perceptual prior (a variant of Griffiths and Kalish[49] model proposed in Langlois et al.[50]).

First, we perform averaging (smoothing) of the raw preference data from Experiment 6 (Figure S2), which consists of 15,000 subjective ratings $U(I_n)$ on 15,000 intervals (I_n) sampled uniformly in the range of [-15 to 15] semitones. To average the data, we computed for every interval i the smoothed function $U'(i)$ using the following formula, which intuitively corresponds to smoothening the data with a Kernel width of B semitones:

$$U'(i) = \sum_n U(I_n) w_n(i) \quad \text{(Equation 9)}$$

Where $w_n(i)$ are smoothing kernels:

$$w_n(i) = C(i) \cdot exp\left((-i - I_n)^2 \Big/ 2B \right) \quad \text{(Equation 10)}$$

$C(i)$ is a normalization constant so that $\sum_n w_n(i) = 1$, and B is the kernel width, set to 0.25 semitones.

This subjective pleasantness function $U'(i)$ can be "translated" to a prior, via a normative model described in Harrison et al.[47]:

$$p(i) = C \cdot exp(\gamma U'(i)) \quad \text{(Equation 11)}$$

where C is a normalization constant so that $\int p(i) di = 1$, and γ is a constant that determines the "peakiness" of preference (how sensitive the prior is to changes in the subjective utility).

We now use a normative Bayesian model that, given a prior, predicts participants' responses. (see detailed justification and assumptions in Langlois et al.,[50] and another application of this model in Jacoby and McDermott[23]). According to this process, the stimulus interval at time t, i_t, is encoded with some sensory noise resulting in an internal representation i'_t. The participant is assumed to decode this internal representation and form a response interval i_{t+1}, which then becomes the input of a new iteration. For simplicity, this model also assumes no production noise. This process can be described with the following Markov chain:

$$\ldots \rightarrow I_t \rightarrow I'_t \rightarrow I_{t+1} \rightarrow \ldots \quad \text{(Equation 12)}$$

Similar to signal detection theory, we assume that $p(I'_t|I_t)$ is an unbiased Gaussian noise with standard deviation of σ (a free parameter):

$$p(I'_t|I_t) \sim N(I_t, \sigma^2) \quad \text{(Equation 13)}$$

We also assume a Bayesian decoding, which means that the response is sampled from the Bayesian inversion:

$$p(I_{t+1} = i_{t+1}|I'_t = i'_t) = p(I_t = i_{t+1}|I'_t = i'_t) = \frac{p(I'_t = i'_t|I_t = i_{t+1}) p(I_t = i_{t+1})}{\int p(I'_t = i'_t|I_t = i_t) p(I_t = i_t) di_t} \quad \text{(Equation 14)}$$

Note that this formula depends on the prior $p(i)$. The only free parameters of the model are σ (noise magnitude, originating from the Gaussian likelihood $p(I'_t|I_t)$) and γ (prior sharpness). In the simulations, we analytically computed the conditional distributions using a grid of width 0.1 semitones, and then sampled points from the model according to Equations 12, 13, and 14. Figure 4H shows the

Please cite this article in press as: Anglada-Tort et al., Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.02.070

CellPress
OPEN ACCESS

aggregated results of the model in the last 3 generations of the simulation (Figure S2 shows an example of the model across generations). The code for the simulations is part of the OSF repository associated with this paper: https://doi.org/10.17605/OSF.IO/UANGD

Combined model

We also explored a combined model. In this model we first used the preference model to generate an expected target interval ($P(I_S)$). We then added production noise just like in the interval-size model, resulting in the following equation:

$$I_R = P(I_S) + b(P(I_S)) + n \quad \text{(Equation 15)}$$

Note that this model has 4 degrees of freedom: the noise magnitude (σ') and the bias scaling (c) from the preference model, and the perceptual noise magnitude (σ) and the prior sharpness (γ). Figure 4I shows the aggregated results of the combined model in the last 3 generations of the simulation (Figure S2 shows an example of the model across generations).The code for the simulations is part of the OSF repository associated with this paper: https://doi.org/10.17605/OSF.IO/UANGD

Model performance and parameter optimization

The interval-size model and preference model had each two free parameters (σ' and c for the interval-size model, and σ and γ for the preference model), which were optimized empirically based on the data from the last 3 iterations of the singing data of Experiment 5. In the case of the combined model, we had 4 parameters, and thus used a lower resolution grid search on a restricted area of the large search space where, based an exploratory analysis indicated that (a) both models contributed to the results so the parameters in that area are not degenerated and (b) we could obtain relative high scores.

We found the best-performing parameters by performing a grid search over the parameter space as explained above and computing the performance for each possible parameter combination. To compute the performance of the model, we bootstrapped 1,000 datasets by sampling chains with replacements from the initially randomized intervals of the singing data (generation 0), and then computed the model predictions for generations 1-10 (see an example of the simulated data across generations in Figure S2). Finally, we computed the marginal across all intervals with KDE with a bandwidth of 0.25 semitones and compared the average KDE of the simulations with singing data using JSD (see comparing distributions).

Model interpretation

Our models are not perfect in separating perception and production, but they provide a useful complementary approach in addition to the behavioral data. In the Interval-size model, we used singing data from the first generation of the singing experiment to model a production bias based on interval size and direction. This data includes a production bias but may also include a perceptual bias (or a mixture of both). However, we only capture a simple form of this data (using the 7th order polynomial), so we can describe the distribution using a few parameters. More importantly, we only take the data from the first generation, and thus the projections for all other iterations including the final distribution are not circular. An alternative approach would require using different theoretic constraints (e.g., vocal range), but the exact way in which these constraints produce biases is not known. In the preference model, we model melodic preferences using data from a listening experiment collecting preference ratings for individual melodic intervals. This task is very different to the production task involved in iterated singing. Despite this, our models provide a useful "first-order approximation" to help interpret the precise contribution of perceptual and production biases on melodic transmission. Follow-up work should extend our approach by testing more refined models of perception and production in singing.