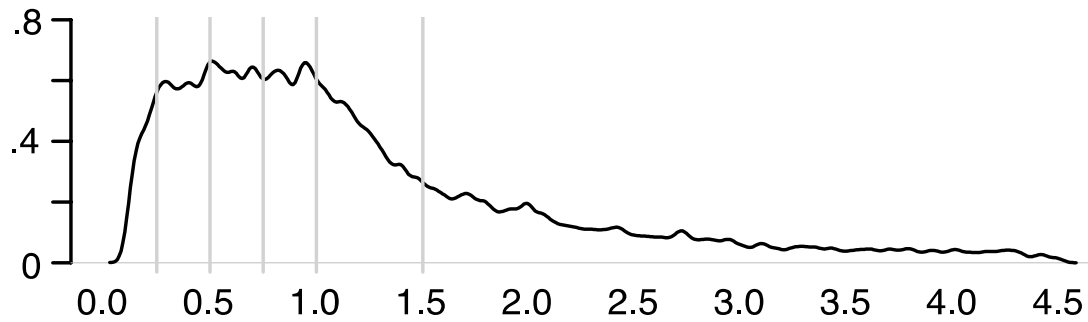


Musical evolution in the lab exhibits rhythmic universals

Andrea Ravignani, Tania Delgado, Simon Kirby

Supplementary Figures



Supplementary Figure 1: Non-adjacent IOIs ratios distribution. For each pattern, we calculated the ratios between non-adjacent IOIs in that pattern (IOI_1/IOI_3 , IOI_1/IOI_4 , ..., IOI_3/IOI_5 , ...), pooling frequency distributions across all 32 patterns from all last generations' participants. This histogram excludes data plotted in Figure 4a. The vertical lines denote the 1:2, 1:1, 1:3, 2:1, and 3:2 ratios.

Supplementary Tables

Supplementary Table 1. Results of the K-clustering algorithm. Each row reports the three centroids (columns 2, 3 and 4), or their median across chains, and their integer ratios (approximated to one decimal digit, last 3 columns) found in the last generation of that particular chain.

Chain\Centroid (ms)	1st	2nd	3rd	2nd/1st	3rd/1st	3rd/2nd
C1	203	493	904	12:5	22:5	9:5 $\approx$ 2
C2	204	398	786	19:10 $\approx$ 2	19:5 $\approx$ 4	2:1=2
C3	202	355	764	9:5	19:5 $\approx$ 4	21:10 $\approx$ 2
C4	253	570	1125	23:10	9:2 $\approx$ 4	2:1=2
C5	177	436	665	5:2=2.5	19:5 $\approx$ 4	3:2=1.5
C6	224	435	690	19:10 $\approx$ 2	31:10 $\approx$ 3	8:5
median	203	435	775			

Supplementary Table 2. Kolmogorov-Smirnov test between frequency of duration for all chains returns small p-values. The distribution of durations is significantly different between the last generations of each chain. These values remain significant even after a conservative Bonferroni correction is applied (6 hypotheses per chain tested at $\alpha=0.05$, each hypothesis tested at $\alpha=0.008$).

	C1	C2	C3	C4	C5
C2	D = 0.2951 p-value = 2.998e-10				
C3	D = 0.2956 p-value = 1.98e-11	D = 0.1865 p-value = 1.718e-04			
C4	D = 0.139 p-value = 7.819e-3	D = 0.2691 p-value = 8.362e-09	D = 0.2683 p-value = 9.305e-10		
C5	D = 0.1665 p-value = 5.778e-4	D = 0.2728 p-value = 3.026e-09	D = 0.2645 p-value = 9.791e-10	D = 0.2934 p-value = 9.5e-12	
C6	D = 0.1926 p-value = 8.416e-06	D = 0.2546 p-value = 4.686e-09	D = 0.2683 p-value = 2.368e-11	D = 0.2156 p-value = 2.197e-07	D = 0.2139 p-value = 1.722e-07

Supplementary Table 3. There is no significant difference (t-tests in table) between the post-test questionnaire responses between the first and final generation of all chains.

	First versus Final Generation
Question 1	t = 0.2644; p-value = 0.7988
Question 2	t = -0.7122; p-value = 0.4995
Question 3	t = -1.409; p-value = 0.1899
Question 4	t = -0.6694; p-value = 0.5191

Supplementary Notes

Increase in systematic structure, learnability and chain-specificity. We find an increase in structural similarities and combinatorial structure over generations (Figure 3), independently from inclusion (Page's trend test; $L=1584.0$, $m=6$, $n=9$, $p<0.001$) or exclusion ($L=1110.0$, $m=6$, $n=8$, $p<0.001$) of generation 0 (i.e. random patterns fed to the 1st participant). Reproduction errors decrease over generations, when both the initial (random) set of patterns is included (Page trend test; $L=833$, $m=6$, $n=8$, $p<0.0001$) and excluded ($L=617$, $m=6$, $n=7$, $p=0.0247$). Chains significantly diverge over generations, no matter whether the initial random generation is included ($L=1558.0$, $m=6$, $n=9$, $p<0.001$) or excluded ($L=1112.0$, $m=6$, $n=8$, $p<0.001$).

Centroids and centroids' ratios. There are several measures and methods to decide on the optimal number of clusters, although none is universally accepted or recognized as optimal. We decided to adopt the simplest, and easiest to explain, approach: the Elbow method [Thorndike (1953). Who Belongs in the Family? *Psychometrika*. 18 (4): 267–276.]. The logic behind this method is: increasing the number of clusters will keep increasing the amount of explained variance (or decreasing the sum of squares) in quite big steps until a certain point; after having reached an optimal number of clusters, the amount of explained variance will only keep increasing marginally, suggesting it is more parsimonious to stop adding clusters. This can be visualized with an Elbow plot for each distributions for which clusters are sought. Using this method, we could show that $k=3$ is the optimal cluster number for all chains.

We tested the null hypothesis that, apart from an effect of centroid number, 1st vs. 2nd vs. 3rd (blocks) the IOI of centroids is the same in each of the chains. A Friedman rank sum test could not reject the null hypothesis (Chi-square=9.857, $N=18$, $df=5$, $p=.079$). This provides some evidence that clusters are the same across chains, although the relative low p-value and degrees of freedom advise caution in interpreting this result.

We tested the hypothesis that the most accented beats occur within binary, ternary and quaternary metrical subdivisions. This hypothesis is not supported ($J=.064$, $p'=0.126$). For non-adjacent ratios (median=.967), only one peak at 1:2 was found (Supplementary Figure 1). However matches in these ratios seem attributable to chance ($J=.064$ $p'=0.743$).

Ratios between adjacent intervals, non-adjacent intervals and intervals between strong beats. Final-generation empirical ratios are not random: KS test between empirical ratios and simulated uniform ratio distribution (100 million data points):

- adjacent: $D=0.085876218644067803$, $pvalue=2.4059408469009131e-13$

- non-adjacent: $D=0.082737297571951349$, $pvalue=1.1048954684067722e-60$
- strong beats: $D=0.088110146122931457$, $pvalue=0.0025929457744099291$

Number of hits. An additional consideration on the quantitative analysis of structure is the length, or number of hits contained by a single pattern. Our experiment program did not reject patterns that were shorter than the initial training patterns, thus any decreases in length would be transmitted to the next generation. A Page's L trend test shows that length of the patterns significantly decreases by the final generation across chains, with a mean of approximately two hits lost per pattern (Generation 0 included: $L=1181$, $m=6$, $n=9$, $p<0.0006$; Generation 0 excluded: $L=1080$, $m=6$, $n=8$, $p<0.003$). The decrease in the number of hits may reflect that participants are unable to remember the entire pattern and fail to reproduce the final hits. A loss in the number of hits could lead to patterns becoming easier to learn over time, in addition to structural regularities in timing.

Post-test Questionnaire. Participants were asked to complete a questionnaire after finishing the behavioral task. The responses to each question were reported on a continuous scale, from 'not at all' to 'extremely'. The questions were: (i) How difficult was the task? (ii) How similar were the patterns to each other? (iii) How musical did the patterns sound? (iv) How familiar did the patterns sound?

Responses were recorded on a continuous scale that was measured in centimeters, which was then normalized for analysis. Scores closer to the value 0 represent responses that were closer to "not at all" and scores closer to the value 1 represent responses closer to "extremely" on a continuous scale. The results of the post-test questionnaire show that participants thought the drumming task was difficult, with an average of 0.69 rating. In accordance with the self-similarity results, participants rated that the patterns were not very similar to each other, with an average score of 0.52. Overall, participants judged the patterns as not musical with an average score of 0.39. The participants also rated the patterns as not sounding familiar with an average rating of 0.41, agreeing with the responses of the patterns not being musical. Between the initial generation and the final generation, there was no significant difference in how the participants responded to the difficulty and musicality of the task (Supplementary Table 3), despite improvement in performance across generations.

The post-test questionnaire revealed that participants thought the task was very difficult. The initial patterns of the experiment were intentionally designed to be difficult with twelve hits of drum sounds that randomly vary along velocity and inter onset duration. Throughout the experiment, some participants used techniques in attempt to more accurately recreate the drum patterns they heard. The varying learning techniques included: tapping along with the pattern using their left hand, mimicking actions needed to create drum sounds using the drumstick, swaying their body, bobbing their head,

mouthed the drum pattern, and tapping their feet. The techniques always followed the drum hits of the pattern they heard. Learners never tapped along with a tempo, which would be difficult to do, given the short length of the patterns and the need for quick memorization. Despite increased accuracy in performance over the course of generations, learners in the final generation of a chain still rated the task as difficult. Learners also rated that the patterns were not very musical across all generations of the chains, although there is evidence of structured patterns emerging within the later generations.

Supplementary Methods

The initial and independent drum patterns were randomly generated by script using the RtMidi library through a Python programming language wrapper and GarageBand software from Apple Inc. The RtMidi library allows for simple processing of MIDI information into MIDI sequencer software such as GarageBand. In GarageBand, the SoCal drum kit was selected to produce drum sounds from the MIDI notes. MIDI note number 40 at pitch E1 was used for the snare drum sound and MIDI note number 31 at pitch G0 was used for the cymbal sound.