Published in final edited form as:

Nature. 2020 June ; 582(7813): 539–544. doi:10.1038/s41586-020-2397-3.

Hidden neural states underlie canary song syntax.

Yarden Cohen[*,1], **Jun Shen**[2], **Dawit Semu**[1], **Daniel P. Leman**[1], **William A. Liberti III**[1,3], **L. Nathan Perkins**[1], **Derek C. Liberti**[4,5], **Darrell N. Kotton**[4,5], **Timothy J. Gardner**[1,6]

[1]Department of Biology, Boston University, Boston, MA 02115, USA

[2]Boston University Center for Systems Neuroscience, Boston, MA 02115, USA

[3]Department of Electrical Engineering and Computer Science, UC Berkeley, Berkeley, CA 94720, USA

[4]Center for Regenerative Medicine of Boston University and Boston Medical Center, Boston, MA 02118, USA

[5]The Pulmonary Center and Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA

[6]Phil and Penny Knight Campus for Accelerating Scientific Impact, University of Oregon, Eugene, OR 97403-6231.

[*]Corresponding: yardencsmail@gmail.com.

Materials & Correspondence

Correspondence and material requests should be addressed to Yarden Cohen or Tim Gardner.

Author contributions

Y.C. and T.J.G. conceived and designed the study. W.A.L.III. designed miniaturized microscopes and tether commutators and consulted on surgical procedures. L.N.P. created the video acquisition software. D.C.L. and D.K. produced lentivirus. Y.C. and J.S. designed surgery procedures. Y.C., J.S., and D.S. performed animal surgeries. Y.C and D.P.L. built the experimental setup. Y.C. and D.S. performed histology and immunohistochemistry. Y.C. and J.S. gathered the data. Y.C. designed and wrote machine-learning audio segmentation and annotation algorithm. Y.C. analyzed the data. Y.C., W.A.L.III, L.N.P., and T.J.G. wrote the manuscript.

Competing interests

The authors declare no competing interests

Data availability

Data can be found at figshare (https://figshare.com/) at 10.6084/m9.figshare.12006657.

Code availability

All custom-made code in this manuscript is publicly available in Github repositories. URLs are provided in the relevant methods descriptions.

Supplementary Information (SI)

Appendix A - Supplementary analysis of overlapping ROIs

This appendix estimates the number of independent sources of fluorescence signal among regions of interests (ROIs) defined in the main text. The appendix shows that unifying overlapping ROIs as one source persisting across days does not change the finding in the main manuscript.

Appendix B - Non-parametric statistical analysis for results in main manuscript

This appendix repeats the manuscript's analyses with the non-parametric 1-way ANOVA (Kruskal Wallis). These analyses recapitulated all the findings in the manuscript.

Supplementary videos 1–7

Video frames show stacks of confocal microscopy sections (3μm thick) that were used to test the specificity of GCaMP expression to excitatory neurons (methods). In all videos, GCaMP is stained in green and the inhibitory neuron markers (CB,CR,PV annotated in the file names) are stained in blue.

Supplementary video 8

The results of the CNMFE[49] algorithm that was used to de-noise the fluorescence videos to visualize context dependent neurons in Figure 4a.

Abstract

Coordinated skills such as speech or dance involve sequences of actions that follow syntactic rules in which transitions between elements depend on the identity and order of past actions. Canary songs are comprised of repeated syllables, called phrases, and the ordering of these phrases follows long-range rules[1], where the choice of what to sing depends on song structure many seconds prior. The neural substrates that support these long-range correlations are unknown. Using miniature head-mounted microscopes and cell-type-specific genetic tools, we observed neural activity in the premotor nucleus HVC[2–4] as canaries explore various phrase sequences in their repertoire. We find neurons that encode past transitions, extending over 4 phrases and spanning up to 4 seconds and 40 syllables. These neurons preferentially encode past actions rather than future actions, can reflect more than a single song history, and occur mostly during the rare phrases that involve history-dependent transitions in song. These findings demonstrate that HVC dynamics includes “hidden states” not reflected in ongoing behavior – states that carry information about prior actions. These states provide a possible substrate to control syntax transitions governed by long-range rules.

Canary songs, like many flexible behaviors contain complex transitions - points where the next action depends on memory of choices made several steps in the past. Songs are composed of syllables produced in trilled repetitions known as phrases (Figure 1a) that are about one-second-long and sang in sequences, typically 20–40 seconds long. The order of phrases in a song exhibit long-range syntax rules[1]. Specifically, phrase transitions following about 15% of the phrase types depend on the preceding sequence of 2–5 phrases. These long-range correlations extend over dozens of syllables, spanning time intervals of several seconds (Figure 1b,c).

In premotor brain regions, neural activity supporting long-range complex transitions will reflect context information as redundant representations of ongoing behavior[5–8]. Such representations, referred to here as ‘hidden neural states’, are predicted in models of memory-guided behavior control[9], but are challenging to observe during unconstrained motion in mammals[10–17] or in songbirds with simple syntax rules[18].

Like motor control in many vertebrate species, canary song is governed by a ‘cortico-thalamic loop’[19–21] that includes the premotor nucleus HVC[2–4]. In stereotyped songs of zebra finches, HVC projection neurons (PNs) produce stereotyped bursts of activity time-locked to song[3]. These cells drive motor outputs or relay timing references to the basal ganglia[22]. In the more variable syllable sequences of Bengalese finches, some projection neurons fire differently depending on the neighboring syllables[18], supporting sequence generation models that include hidden states[9]. However, the time-frame of the song-sequence neural correlations are relatively short (roughly 100ms). In contrast, correlations in human behavior can extend for tens of seconds and beyond and support long-range syntax rules. At present it is not known if redundant premotor representations in songbirds can support working memory for syntax control over timescales longer than 100ms.

To further dissect the mechanisms of working memory for song we used custom head mounted miniature microscopes to record HVC PNs during the song production in freely

moving canaries (serinus canaria) (Figure 2b). Although PNs form distinct projection-target-specific subtypes, the imaging method does not distinguish these populations and we report results for this mixed population as a whole. These experiments reveal a previously un-described pattern of neural dynamics that can support structured, context-dependent song transitions and validate predictions of long-range syntax generation by hidden neural states[9,23] in a complex vocal learner.

Complex transitions in a subset of song phrases

Inspired by technological advances in human speech recognition[24], we developed a song segmentation and annotation algorithm that automated working with large datasets (>5000 songs; Extended Data Fig. 1a, methods). The birds' repertoire included 24–37 different syllables with typical durations between 10–350 msec. The average number of syllable repeats per phrase type ranged from 1 to 38 with extreme cases of individual phrases exceeding 10 seconds and 120 syllables (Extended Data Fig. 1c–g). Transitions between phrases could be completely deterministic, where one phrase type always followed another, or flexible, where multiple phrase types followed a given phrase (sonograms in Figure 1a, stack in Figure 1b). In very rare cases, transitions contained an aberrant syllable that could not be stably classified (c.f. Extended Data Fig. 2g–i), and all data was visually proofed. (Extended Data Fig. 2 and Extended Data Fig. 1b illustrate the reliable annotation of phrase sequences and syllable repertoires).

As shown in another strain of canaries[1], we found that a small subset of phrase types precede 'complex' transitions - behavioral transitions that depend on the multi-step context of preceding phrases. Specifically, the probability of transition outcomes can change by almost an order of magnitude depending on the identity of the 3 preceding phrases (Figure 1b). Such song context dependence is captured by a 3rd order Markov chain. Extended Data Fig. 1i shows the significant long-range context-dependent transitions for two birds.

HVC PNs encode long range syntax

To characterize the neural activity supporting complex transitions, we imaged neurons that expressed the genetically-encoded calcium indicator GCaMP6f in freely-behaving adult male canaries (N=3, age>1yr, recording in left hemisphere HVC[2]). The indicator is selectively-expressed in projection neurons and allows recording neural activity via fluorescence dynamics extracted from annotated regions of interest (ROIs, Extended Data Fig. 4, methods). In our dataset, 95% of all phrases are trills of multiple syllables and only 6.1% of those are shorter than the decay time constant of the calcium indicator (400 msec[25], Extended Data Fig. 1h). As in finches, we found that HVC projection neuron activity in canaries was sparse in time[3,18]. Out of N = 2010 daily-annotated ROIs (35 +/−15, mean +/−SD ROIs per animal per day), about 90% were selectively-active in just one or two phrase types (Figure 2a,c, Extended Data Fig. 5a). This, combined with the long phrase duration (Extended Data Fig. 1f,h), allowed us to examine song-context dependence of neural activity using GCaMP6f. In the following, recordings in different days are treated separately. This approach overestimates the number of independent neurons we imaged but avoids analysis

biases and stability concerns. Under the more conservative assumption of sources persisting across days, in SI Appendix A we still estimate 1057 independent sources in our dataset.

Examining the patterns of phrase-locked activity, we identified signals that change depending on song context. For example, ROIs showing weaker or no activity in one song context while demonstrating strong activity in another song context (Figure 2a). Importantly, we found that this context-dependent activity is strongly influenced by the identity of non-adjacent phrases. For example, Figure 2d shows the de-noised fluorescence signal raster from a ROI, locked to the phrase type marked in pink, displaying a dramatic variation in activity ($(\Delta f/f_0)_{denoised}$, methods) depending on the 2nd phrase in the sequence's past – a 2nd order correlation. This sequence preference is quantified by integrating the ROI-averaged signal (Extended Data Fig. 5b,c, 1-way ANOVA, $F(5,35)=18.3$, $p<1e\text{-}8$. 1-way ANOVA evaluates the null hypothesis of no activity variation with phrase identity for all sequence-correlated ROIs in this manuscript). We found ROIs with signals that relate to the identity of past and future non-adjacent phrases in all 3 birds (Extended Data Fig. 5). Across all animals, 21.2% of the daily annotated ROIs showed sequence correlations extending beyond the current active syllable. 18.1% had 1st order correlations where activity during one phrase depends on the identity of an adjacent phrase, and 5.6% had ≥2nd order relations (Extended Data Fig. 5d).

These sequence dependencies could potentially be explained by other factors inherent to the song that may be more predictive of phrase sequence than HVC activity. For example, transition probabilities following a given phrase could potentially depend on the phrase duration[1], on the onset and offset timing of previous phrases, and on the global time since the start of the song – implicating processes such as neuromodulator tone, temperature buildup, or slow adaptation to auditory feedback[26–31] (c.f. Extended Data Fig. 6a–g). To rule out these explanations, we used multivariate linear regression and repeated the tests for sequence-correlated neural activity after discounting the effect of these duration and timing variables on the neural signals. We found that 32.8% (39/119 from 3 birds) of ≥2nd order relations and 52.7% (147/279 from 3 birds) of the 1st order relations remain significant (Extended Data Fig. 5c,6h).

The sequence-correlated ROIs tend to reflect past events more often than future events. Out of N=398 significant phrase sequence-neural activity correlations, 62.3% reflect preceding phrase identities (binomial z-test rejects the hypothesis of 50%, $z=6.94$, $p<1e\text{-}11$). This bias is also found separately in 1st or higher order correlations (Figure 2e, 60.2% and 67.2% respectively. Both percentages are significantly larger than 50%, binomial z-test, $z=4.82,5.31$ $p<1e\text{-}6$, $p<1e\text{-}6$ respectively and oppose the bias of 44.6% and 43.1% 1st and 2nd order correlations expected to reflect past events from behavior statistics alone, $p<1e\text{-}7$, binomial tests) and persists if we consider ROIs that overlap in footprint and sequence correlation across days as the same source (SI Appendix A). Apart from being more numerous, past correlations also tend to be stronger than future correlations (Extended Data Fig. 6i, significantly larger mean fraction explained variance (η^2) in past correlation, bootstrap comparison rejects the null hypothesis of equal means, $p<1e\text{-}6$, $p=0.001$ for 1st and higher order correlations respectively).

These findings suggest that for a subset of HVC neurons, calcium signals are not just related to present motor actions, but convey the context of past events across multiple syllables.

HVC PNs also encode within-phrase timing

HVC projection neurons have been recorded in Bengalese finches and in swamp sparrows, two species that also sing strings of syllable repeats. In swamp sparrows, examples of basal ganglia projecting HVC neurons exhibited stereotyped syllable-locked firing for each syllable in a repeated sequence[32]. In Bengalese, the same pattern was described for some cells as well as ramping syllable-locked spike bursts that increase or decrease in spike number over the course of a phrase[18]. In our dataset, a small subset of ROIs is consistent with fixed syllable-locked neural activity (Figure 2c, Extended Data Fig. 7a,c). More commonly, the activity is restricted to a brief period of time within a phrase as in Figure 2d, not time locked to each syllable within the phrase. Examining all sequence-correlated ROIs we found that 91% are active for time-intervals shorter than the phrase with peak timing and onset timing that can be found at all times in the phrase (Figure 3, Extended Data Fig. 7b,c,e also showing some transients could be explained by ramping syllable-locked spike bursts). Together, these findings indicate that the majority of neurons recorded here contain information about timing within a phrase, not just syllable identity.

PNs carrying long range information

Long-range syntax rules imply that a memory of previous elements sung influences future syllable choice. The HVC activity described here provides a clue for a possible mechanism. For example, during a fixed sequence of four phrases, we found ROIs that carry forward information about the identity of the first phrase during each subsequent phrase. (Figure 4a,b, Extended Data Fig. 8a, 1-way ANOVA showing significant modulation of neural activity with the identity of the past phrase). In this example, the ROIs that reflect long-range information continue to do so even if the final phrase in the sequence is replaced by the end of song, suggesting that their activity reflects prior song context rather than some upcoming future syllable choice. (Extended Data Fig. 8b, 1-way ANOVA, $F(5,10)=36.14$ and 2.79, $p<5e\text{-}6$ and $p<0.08$ for ROIs 50 and 36, when replacing the last phrase with the end of song). This example suggests that a chain of neurons reflecting “hidden states” or information about past choices could provide the necessary working memory to implement long-range transition rules.

HVC PNs active in complex transitions

The phrases in Figure 4 are phrase types that lead to complex transitions or directly follow them (in Figure 1). If HVC neurons with context-selective activity are driving long-range syntax rules, then they should represent song context information predominately around complex behavior transitions, when such information is needed to bias transition probabilities. Accordingly, at the population level, we found more sequence-correlated ROIs around complex transitions; about 70% of sequence-correlated ROIs were found during the rare phrase types that immediately precede or follow complex transitions (Figure 4c, 76%(65%) for 1^{st}($\geq 2^{nd}$) order). Both percentages are larger than the 27%(22%) expected

from uniform distribution of sequence-correlations in all phrases (binomial test, p<1e-10, Extended Data Fig. 8c–f) and persist if we consider ROIs that overlap in footprint and sequence correlation across days as the same source (SI Appendix A). Separating the influence of past context and future action on the neural activity we find that, in complex transitions, ROIs predominately represent the identity of the preceding phrase (Extended Data Fig. 8g,h, multi-way ANOVA test and Tukey's post-hoc analysis showing that the preceding phrase identity significantly affects the neural activity more than twice more often than the following phrase identity. Binomial z-test rejects the null hypothesis of equal groups. Z=6.45, p<1e-10). This bias does not occur outside of complex transitions (Extended Data Fig. 8i, binomial z-test, Z=1.06, p>0.1). This finding suggests that neural coding for past context is enriched during transitions that require this context information.

Synergistic activity predicts complex behavior

For the ROIs with first and ≥2nd order sequence correlations, 19% and 14% respectively were active in several preceding phrase contexts, whereas 44% and 48% preferred just one out of several past contexts (Extended Data Fig. 9). Neurons responding in multiple contexts can complement each other to provide additional information about song history (Figure 4 ROIs 21,45,50, Figure 2d). Extended Data Fig. 10a, shows four ROIs that were jointly active during a single phrase type. One ROI was active in a single context (ROI 10) and the other three were active in multiple contexts. The phrase during which these ROIs were recorded precedes a complex transition and, in this example, the behavior alone (prior phrase type) poorly predicts the transition outcome (right bar in Extended Data Fig. 10b, 0.08 out of 1, bootstrapped normalized mutual information estimate, methods). However, looking at multiple ROIs together we found that the network holds significantly more information about the past and future phrase types (Extended Data Fig. 10b, 0.42,0.33, bootstrapped z-test rejects the null hypothesis of equal means, z=8.95, p<1e-15). This increase exceeds the most informative individual ROIs (0.33, 0.21, Bootstrapped z-test rejects the null hypothesis of equal means z=2.26, p<0.015 and z=5.7, p<1e-8 respectively), suggesting synergy of the complementing activity patterns. Furthermore, in this example the network holds more information about the past than the future (Extended Data Fig. 10b–d, bootstrapped z-test, z=4.32, p<1e-5), suggesting that information is lost during the complex transition.

Taken together, these findings demonstrate that neural activity in canary HVC carries long range song context information. These "hidden states" relate primarily (c.f. Extended Data Fig. 3) to past or future song and contain the information necessary to drive complex, context-dependent phrase transitions.

Discussion

Motor skills with long-range sequence dependencies are common in complex behaviors, with speech the richest example. In general, the neural mechanisms underlying long-range motor sequence dependencies are unknown. Here we show that context sensitive activity in HVC projection neurons can support the long-range order in canary song sequences[1]. Specifically, we find projection neurons whose activity is contingent on phrases up to four steps in the past and projection neurons predicting phrases two steps into the future. Cells

with this higher order behavior tend to be active during complex behavioral transitions – times when the song behavior requires high level information about the sequence context. A key next step will be to further subdivide the activity reported here, determining which projection neuron classes in HVC carry the long-range information.

The HVC activity described here resembles the many-to-one relation between neural activity and behavior states[9,23,27,33] proposed in some models to relay information across time. In this respect, our findings expand on a prior study in Bengalese finches[18], that showed HVC projection neurons whose activity depended not just on the current syllable type but also the prior syllable type. This history extended just to the most recent syllable transition, over a time frame of roughly 100ms.

In the canary HVC neurons observed here, the time frame extends over multiple phrases - several seconds. This longer time frame rules out explanations based on short term biophysical processes including short-term calcium dynamics, synaptic plasticity[34], channel dynamics[35] supporting auditory integration[36], sensory-motor delay, and adaptation to auditory inputs[27] that could span a smaller 50–250msec time frame. Unlike syllable-locked neural activity reported in Bengalese finches[18], the onset of hidden state activity in canaries is not restricted to phrase edges. Rather, the activity recorded here suggests that parallel chains of sparse neural activity propagate in the song system during a given phrase and that distinct populations of neurons can sequentially encode the same syllable type – a many to one mapping of neural sequences onto syllable types predicted by a prominent statistical model of birdsong[9].

There are clues that HVC does not contain all the information required to select a phrase transition – since more neurons correlate to the sequence's past than to its future it is possible that sequence information in HVC is lost, perhaps due to neuronal noise that adds stochasticity to transitions. The source of residual stochasticity in HVC could be intrinsic to the dynamics of HVC – resembling the "noise" terms commonly added in sequence generating models[37–39] or may enter downstream as well-documented noise in the basal ganglia outputs[40] also converge on pre-motor cortical areas downstream of HVC and may impact phrase transitions.

The study of neural dynamics during flexible transitions in canaries may provide a tractable model for studying stochastic cognitive functions – mechanisms in working memory and sensory-motor integration that remain extremely challenging to quantify in most spontaneous behaviors in mammals. Finally, we note that recent dramatic progress in speech recognition algorithms have employed recurrent neural networks with several architectures designed to capture sequence dependencies with hidden states. Examples include LSTM[41], hierarchical time scales[42], hidden memory relations[43], and attention networks[44]. It is possible that machine learning models will help to interpret the complex dynamics of the song system, and help inform new models of many to one, history dependent mappings between brain state and behavior[23].

Materials and Methods

Ethics declaration

All procedures were approved by the Institutional Animal Care and Use Committee of Boston University (protocol numbers 14–028 and 14–029).

Subjects

Imaging data were collected from $n = 3$ adult male canaries. Birds were individually housed for the entire duration of the experiment and kept on a light–dark cycle matching the daylight cycle in Boston (42.3601 N) with unlimited access to food and water. The sample sizes in this study are similar to sample sizes used in the field. The birds were not used in any other experiments. This study did not include experimental groups and did not require blinding or randomization.

Surgical procedures

Anesthesia and analgesia—Prior to anesthetizing the birds, they were injected with meloxicam (intramuscular, 0.5mg/kg) and deprived of food and water for a minimum of 30 minutes. Birds were anesthetized with 4% isoflurane and maintained at 1–2% for the course of the surgery. Prior to skin incision, bupivacaine (4 mg/kg in sterile saline) injected subcutaneously (volume 0.1–0.2 mL). Meloxicam was also administered for 3 days after surgery.

Stereotactic coordinates—The head was held in a previously described, small animal stereotactic instrument[45]. To increase anatomical accuracy and ease of access, we deviated from the published atlas coordinates[45] and adapted the head angle reference to a commonly used forehead landmark parallel to the horizontal plane. The outer bone leaflet above the prominent λ sinus was removed and the medial (positive = right) and anterior (positive) coordinates are measured from that point. The depth is measured from the brain's dura surface. The following coordinates were used (multiple values indicate multiple injections):

HVC: +65°, −2.5mm ML, 0.12mm AP, 0.15–0.7mm D

RA: +80°, −2.5mm ML, −1.2mm AP, 1.9–3mm D

Area X: +20°, −1.27, −1.3mm ML, 5.65, 5.8mm AP, 2.65–2.95mm D

Angles (°) are measured from the horizontal plane defined above and increase as the head is rotated downward, the mediolateral coordinate (ML) is measured from the midline and increases rightward, the anterior-posterior coordinate (AP) is parallel to the horizontal plane and measured forward from λ, and the Depth (D) is measured from the brain's surface and increase with depth.

HVC demarcation and head anchoring—To target HVC, 50–100nL of the DiI retrograde lipophilic tracer (5mg/ml solution in dimethylformamide, DMF) was injected into the left area X. The outer bone leaflet was removed above area X with a dental drill. The inner bone leaflet was thinned and removed with an ophthalmic scalpel, exposing a hole of

~300 μm diameter. The left area X was injected using a Drummond Nanoject II (Drummond pipette, 23nl/sec, pulses of 2.3nl). In the same surgery, a head anchoring structure was created by curing dental acrylic (Flow-It ALC, Pentron) above the exposed skull and through ~100μm holes in the outer bone leaflet.

Virus injection and lens implants—A lentivirus that was developed for previous work in zebra finches (containing the vector pHAGE-RSV-GCaMP6f; Addgene plasmid #80315) was also used in canaries[46]. The outer skull leaflet above HVC was removed with a dental drill. The inner bone leaflet was thinned and removed with an ophthalmic scalpel, exposing ~1.5–2mm diameter area of the dura. The DiI demarcation of HVC was used to select an area for imaging. The lentivirus was injected in 3–4 locations, at least 0.2mm apart, at a range of depths between 0.5–0.15mm. In total 800–1000nL were injected into the left HVC. After the injection, the dura was removed and the parahippocampus segment above the imaging site was removed with a dura pick and a custom tissue suction nozzle. A relay GRIN lens (Grintech GT-IFRL-100, 0.44 pitch length, 0.47 NA) was immediately positioned on top of the exposed HVC and held in place with Kwik-Sil (WPI). Dental acrylic (Flow-It, Pentron) was used to attach the lens to the head plate and cover the surgery area. The birds were allowed to recover for 1–2 weeks.

Hardware—To image calcium activity in HVC projection neurons during singing, we employed custom, lightweight (~1.8 g), commutable, 3D-printed, single-photon head-mounted fluorescent microscopes that simultaneously record audio and video (Figure 2). These microscopes enabled recording hundreds of songs per day, and all songs were recorded from birds longitudinally in their home cage, without requiring adjustment or removal of the microscope during the imaging period. Birds were imaged for less than 30 min total on each imaging day, and LED activation and video acquisition were triggered on song using previously described methods[46].

Microscope design—We used a custom, open-source microscope developed in the lab[46]. A blue LED produces excitation light (470-nm peak, LUXEON Rebel). A drum lens collects the LED emission, which passes through a 4 mm × 4 mm excitation filter, deflects off a dichroic mirror, and enters the imaging pathway via a 0.25 pitch gradient refractive index (GRIN) objective lens. Fluorescence from the sample returns through the objective, the dichroic, an emission filter, and an achromatic doublet lens that focuses the image onto an analog CMOS sensor with 640 × 480 pixels mounted on a PCB that also integrates a microphone. The frame rate of the camera is 30 Hz, and the field of view is approximately 800 μm × 600 μm. The housing is made of 3D-printed material (Formlabs, black resin). A total of 5 electrical wires run out from the camera: one wire each for camera power, ground, audio, NTSC analog video and LED power. These wires run through a custom flex-PCB interconnect (Rigiflex) up to a custom-built active commutator. The NTSC video signal and analog audio are digitized through a USB frame-grabber. Custom software written in the Swift programming language running on the macOS operating system (version 10.10) leverages native AVFoundation frameworks to communicate with the USB frame-grabber and capture the synchronized audio–video stream. Video and audio are written to disk in MPEG-4 container files with video encoded at full resolution using either H.264 or lossless

MJPEG Open DML codecs and audio encoded using the AAC codec with a 48-kHz sampling rate. All schematics and code can be found online https://github.com/gardner-lab/FinchScope and https://github.com/gardner-lab/video-capture.

Microscope positioning and focusing—Animals were anesthetized and head fixed. The miniaturized microscope was held by a manipulator and positioned above the relay lens. The objective distance above the relay was set such that blood vessels and GCaMP6f expressing cells were in focus. The birds recovered in the recording setup. Within the first couple of weeks, the microscopes were refocused to maximize the number of observable neurons.

Histology—DiI was injected into area X as described above. Three days later, ~800nL lentivirus was injected into HVC using the DiI demarcation. In finches, this virus infected predominately projection neurons. In this project we analyzed neurons with sparse activity that do not match the tonic activity of interneurons in HVC. The virus was injected in four sites, at least 0.2mm apart and at two depths (matching the experiment's procedure). About four weeks later the bird was euthanized (by an intracoelomic injection of 0.2mL 10% Euthasol, Virbac, ANADA #200–071, in saline) and perfused by first running saline and then 4% paraformaldehyde via the heart's left chamber and the contralateral neck vein. The brain was extracted and kept overnight in 4% paraformaldehyde at 4°C.

GCaMP6f expression (Extended Data Fig. 4a): The fixed tissue was sectioned into 70 μm sagittal slices (Vibratome series 1000), placed on microscope slides, and sealed with cover slips and nail polish. Epifluorescence images were taken with Nikon Eclipse Ni-E tabletop microscope.

Expression specificity to excitatory neurons (Extended Data Fig. 4b): The fixed tissue was immersed in 20% and 30% sucrose solutions for two overnights, frozen and sectioned into 30 μm sagittal slices (Cryostat, Leica CM3050S). Following work in zebra finches[47], the slices were stained for calcium binding interneuron markers Calbindin (1:4000, SWANT), Calretinin (1:15000, SWANT), and Parvalbumin (1:1000, SWANT) by overnight incubation with the primary antibody at 4°C and with a secondary antibody (coupled to Alexa Fluor 647) for 2 hours at room temperature. Slices were mounted on microscope slides, and sealed with cover slips and nail polish. A confocal microscope (Nikon C2si) was used to image GCaMP6f and the interneuron markers in 3μm-thick sections through the tissue. The images were inspected for co-stained cells (e.g. SI videos 1–7). The results ruled out any co-expression of GCaMP and Calbindin or Calretinin. We found 2 cells expressing Parvalbumin and GCaMP (SI video 5 shows one example, <0.5% of PV stained cells, <0.01% of GCaMP expressing cells), possibly replicating previous observation of PV expression in HVC projection neurons[47].

Data collection

Song screening—Birds were individually housed in soundproof boxes and recorded for 3–5 days (Audio-Technica AT831B Lavalier Condenser Microphone, M-Audio Octane amplifiers, HDSPe RayDAT sound card and VOS games' Boom Recorder software on a

Mac Pro desktop computer). In-house software was used to detect and save only sound segments that contained vocalizations. These recordings were used to select subjects that are copious singers (≥50 songs per day) and produce at least 10 different syllable types.

Video and audio recording—All data used in this manuscript was acquired between late February and early July – a period during which canaries perform their mating season songs. To avoid over exposure of the fluorescent proteins, data collection was done during the morning hours (from sunrise until about 10am) and the daily accumulated LED-on time rarely exceeded 30 minutes. Audio and video data collection was triggered by the onset of song as previously described[46] with an additional threshold on the spectral entropy that improved detection of song periods dramatically. Data files from the first couple of weeks, a period during which the microscope focusing took place and the birds sang very little, was not used in this manuscript. Additionally, data files from (extremely rare) days in which video files were corrupted because of tethering malfunctions, were not used in this manuscript.

Data Analysis

Video file preprocessing—Software developed in-house was used to load video frames and audio signal to MATLAB (https://github.com/gardner-lab/FinchScope/tree/master/Analysis%20Pipeline/extractmedia) along with the accompanying timestamps. Video frames were interpolated in time and aligned to an average frame rate of 30Hz. Audio samples were aligned and trimmed in sync with the interpolated frame timestamps. To remove out-of-focus bulk fluorescence from the 3-dimensional representation of the video (rows × columns × frames), the background was subtracted from each frame by smoothing it with a 145 pixel-wide circular Gaussian kernel, resulting in 3-dimensional video data, $V(x, y, t)$.

Audio processing—Song syllables were segmented and annotated in a semi-automatic process:

- A set of ~100 songs was manually annotated using a GUI developed in-house (https://github.com/yardencsGitHub/BirdSongBout/tree/master/helpers/GUI). This set was chosen to include all potential syllable types as well as cage noises.
- The manually labeled set was used to train a deep learning algorithm developed in-house (https://github.com/yardencsGitHub/tweetynet).
- The trained algorithm annotated the rest of the data and its results were manually verified and corrected.
- In both the training phase of TweetyNet and the prediction phase for new annotations, data is fed to TweetyNet in segments of 1 second and TweetyNet's output is the most likely label for each 2.7msec time bin in the recording.

Assuring the separation of syllable classes—To make sure that the syllable classes are well separated all the spectrograms of every instance of every syllable, as segmented in the previous section, were zero-padded to the same duration, pooled and divided into two equal sets. For each pair of syllable types, a support vector machine classifier was trained on

half the data (the training set) and its error rate was calculated on the other half (the test set). These results are presented, for example, in Extended Data Fig. 1b.

Testing for within-class context distinction by syllables acoustics—Apart from the clear between-class separation of different syllables for syllables that precede complex transitions we check the within-class distinction between contexts that affect the transition. To do that we use the parameters defined in Wohlgemuth et al., 2010[48] and treat each syllable rendition as a point in an 8 dimensional space of normalized acoustic features. For a pair of syllable groups (different syllables or the same syllable in different contexts) we calculate the discriminability coefficient:

$$d'_{AB} = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{2} + \frac{\sigma_B^2}{2}}},$$

Where, $\mu_A - \mu_B$ is the L_2 distance between the centers of the distributions and σ_A^2, σ_B^2 are the within-group distance variance from the centers. Extended Data Fig. 3 demonstrates that all within-class d' values are smaller than all between-class d' values.

Identifying complex transitions—Complex transitions were identified by the length of the Markov chain, required to describe the outcome probabilities. These dependencies were found using a previously-described algorithm that extract the probabilistic suffix tree (PST[1]) for each transition (https://github.com/jmarkow/pst). Briefly, the tree is a directed graph in which each phrase type is a root node representing the first order (Markov) transition probabilities to downstream phrases, including the end of song. The pie chart in Extended Data Fig. 1i (i) shows such probabilities. Upstream nodes represent higher order Markov chains, 2nd and 3rd in Extended Data Fig. 1i (ii) and (iii) respectively, that are added sequentially if they significantly add information about the transition.

ROI selection, ΔF/F signal extraction and de-noising—Song-containing movies were converted to images by calculating, for each pixel, the maximal value across all frames. These 'maximum projection images' were then similarly used to create a daily maximum projection image and also concatenated to create a video. The daily maximum projection and song-wise maximum projection videos were used to select regions of interest (ROIs), purported single neurons, in which fluorescence fluctuated across songs.

ROIs were never smaller than the expected neuron size, did not overlap, and were restricted to connected shapes, rarely deviating from simple ellipses. Importantly, this selection method did not differentiate between sources of fixed and fluctuating fluorescence. The footprint of each ROI in the video frames was used to extract the time series, $f(t) = \sum_{(x,y) \in ROI} V(x, y, t)$, summing signal from all pixels within that ROI. Then, signals were converted to relative fluorescence changes, $\frac{\Delta f(t)}{f_0} = \frac{f(t) - f_0}{f_0}$ by defining f_0 to be the 0.05 quantile.

The de-noised fluorescence, $(\Delta f/f_0)_{denoised}$, is estimated from the relative fluorescence change using previously published modeling of the calcium concentration dynamics and the added noise process caused by the fluorescence measurement[49]

Seeking ROIs with sequence correlations—Since each ROI was sparsely active in very few phrase types we first sought ROIs that are active during a phrase type and then tested if it shows correlations to preceding or following phrase identities. We used the following 2 steps scheme:

Step 1 – identify ROIs with phrase-type-active signal: Phrase-type-active ROI was defined by requiring signal, $s(t) = \frac{\Delta f(t)}{f_0}$ as defined in the previous section, to be larger and distinct from noise fluctuations (for each ROI and repeats of each phrase type, P):

- The 0.9 quantile, Δff_{90}, is taken as a measure of within-phrase peak values – reducing outliers.
- Irrespective of the phrase boundaries, periods of time when an ROI is active are separated from baseline noise fluctuations by fitting the signal within an ROI, s(t), with a 2-state hidden Markov model with Gaussian emission functions. Specifically, at time t the observable, s(t), is assumed to follow a Gaussian distribution, $\mathbb{N}(\mu_t, \sigma_t)$, that determines the likelihood $p(s(t); \mu_t, \sigma_t)$. The hidden variable, $\boldsymbol{\Theta}_t = (\mu_t, \sigma_t)$, is defined by the mean ($\boldsymbol{\mu} = \mu_1, \mu_2$) and standard deviation ($\boldsymbol{\sigma} = \sigma_1, \sigma_2$) of the Gaussian distributions and follows a 1st order time-independent Markov transition probabilities, $\overleftrightarrow{R} = p(\boldsymbol{\Theta}_{t+1}|\boldsymbol{\Theta}_t)$, a 2×2 matrix of transition probabilities between 2 states ('activity' and 'noise'). To estimate the sequence of states (the hidden process $\boldsymbol{\Theta}$) we maximize the log-likelihood:

$$L\left\{s, \boldsymbol{\Theta}, \overleftrightarrow{R}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right\} = \sum_t \log p(\boldsymbol{\Theta}_t|\boldsymbol{\Theta}_{t-1}) + \sum_t \log p(s(t)|\boldsymbol{\Theta}_t)$$

  In this process, the mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma}$) of the two Gaussian distributions are free parameters.
- We define the phrase-type-occupancy, HMM_P, as the fraction of phrase 'P' repetitions that contained the 'active' state.
- These two activity measures, Δff_{90} and HMM_P, are used to select ROIs to be investigated for sequence correlations. We impose lenient thresholds:
  - $\Delta ff_{90} > 0.1$ (i.e. fluorescence fluctuation is larger than a 10% deviation from baseline)
  - $HMM_P > 0.1$ (i.e. the phrase type carries neural activity in 10% of occurrences or more often). In our data set, this threshold is roughly equivalent to ignoring ROIs active only once or twice during a recording day.

Step 2 – test sequence correlations

- 1st order relationships between the signal integral (summed across time bins in the phrase) and the upstream or downstream phrase identities were tested with 1-way ANOVA.
- The entire set of songs of each bird was used to calculate the 1st order phrase transition probabilities, $P_{ab} = P(a \rightarrow b)$, for all phrases 'a', 'b'.
- 2nd order relationships were tested between the signal integral and the identity of the 2nd upstream (downstream) phrase identity for all intermediate phrase types that preceded (followed) the phrase-in-focus in ≥10% of the repeats (as indicated by the phrase transition matrix)
- Sequence-signal correlations were not investigated if fewer than N = 10 repeats contributed to the test.
- Relations were discarded if the label, leading to the significant ANOVA, contained only one song.
- Data used for ANOVA tests is represented in Extended Data figures by box plots marking the median (center line); upper and lower quartiles (box limits); extreme values (whiskers), and outliers (+ markers).
- The data were not tested for normality prior to performing ANOVA tests for individual neurons with the following reasoning:
  - Statistics textbooks suggest that violating the normality requirement is not expected to have a significant effect. For example, Howell, Statistical Methods in Psychology, Chapman & Hall, 4th Ed writes: "As we have seen, the analysis of variance is based on the assumptions of normality and homogeneity of variance. In practice, however, the analysis of variance is a robust statistical procedure, and the assumptions frequently can be violated with relatively minor effects. This is especially true for the normality assumption. For studies dealing with this problem, see Box (1953, 1954a, 1954b), Boneau (1960), Bradley (1964), and Grissom (2000)."
  - Carrying tests for normality will create a bias in our analyses. Each neuron tested for phrase sequence correlation is recorded in a different number of songs. Testing for normality will bias towards larger numbers of songs and against high-order correlations.
  - Nevertheless, we repeated the analyses in this manuscript with non-parametric one-way analysis of variance (Kruskal - Wallis). While fewer neurons pass the more stringent tests (~15% less), all the results in the manuscript remain the same. We include a summary of the non-parametric statistics as SI Appendix B.

Note: In this procedure, sparsely active ROIs or ROIs active in rare phrase types were not tested for sequence correlation. In the main body we reported that 21.2% of the entire set of

ROIs showed sequence correlation. This percentage includes also ROIs that were not tested for sequence correlations. Out of the ROIs that were tested, about 30% had significant sequence correlations (23% and 10% showed 1st and 2nd order correlations)

Phrase specificity—The fractions of phrase repetitions, during which a ROI is 'active', HMM_P, were also used to calculate the ROIs' phrase specificity (in Figure 2):

- For each ROI, the fraction of activity in repetitions of each phrase was calculated separately.
- These measures were normalized and sorted in descending order.
- The number of phrase types accounting for 90% of the ROI's activity was calculated.

Transition-locked activity onsets (Figure 2e,Extended Data Fig. 7d)—The hidden Markov modeling of neural activity was used to identify signal onsets at transition from the 'noise' to the 'active' states. The phrase transition segment is defined as the time window between the onset of the last syllable in one phrase and the offset of the first syllable in the next phrase. ROIs whose sequence-correlated activity initiates in the phrase transition in the majority of cases were suspected as transition-locked representations. These activity rasters were manually examined and a small number (9) of representations were excluded from population-level statistics because they appeared reliably and exclusively in specific transitions. Signals exclusively-occurring in specific transitions are trivially sequence correlated but simply reflect the ongoing behavior. This exclusion does not change the results in this manuscript.

Controlling for phrase durations and time-in-song confounds—In songs that contain a fixed phrase sequence, as in Figure 2d, we calculated the significance of the relation between $s = \sum_{t \in p}(\Delta f / f_0)_{denoised}$, an integral of the signal during one phrase in the sequence, the target phrase 'p', and the identity of an upstream phrase that changes from song to song using a 1-way ANOVA. This relation can be carried by several confounding variables:

- The duration of the target phrase.
- The relative timing of intermediate phrase edges, between the changing phrase and the target phrase.
- The absolute time-in-song of the target phrase.

In Extended Data Fig. 6h we account for these variables by first calculating the residuals of a multivariate linear regression (a general linear model, or GLM) between those variables and s, and then using 1-way ANOVA to test the relation of the residuals and the upstream or downstream phrase identity.

Comparing numbers of significant sequence correlations to past and future events (Figure 2e)—In Figure 2e we compare the numbers of significant sequence correlations between two groups. Group sizes were converted to fractions and the binomial

comparison z statistic was used to compare those fractions. Generally, the statistic $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ with $\hat{p}_1$, $\hat{p}_2$ the measured fractions of significant correlations in two populations of sizes n_1, n_2 and $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ is tested against the normal distribution null hypothesis of zero mean. The effect size, $\hat{p}_1 - \hat{p}_2$, has the confidence interval $CI = \pm 1.96 \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.

In this comparison there is no bias from the conditions of the statistical test, 1-way ANOVA, used to establish sequence correlations of individual ROIs:

The process of seeking ROIs with sequence correlations (described above) guarantees that tests were not carried in under-sampled conditions because the minimal number of repetitions always exceeded the number of song contexts. In these conditions the ANOVA test is not biased by the number of song contexts, or branching order, in different transitions because the test's significance threshold depends on the number of statistical degrees of freedom that account for the number of contexts. This dependence guarantees that tests with more (or less) song contexts are not more likely to get statistical significance by chance.

Contrasting the strength of sequence correlation to past and future events (Extended Data Fig. 6i)—For 1-way ANOVA tests we estimate the significance of the difference in η^2-statistics (frac. explained variance) calculated in past versus future correlations using the following bootstrapping procedure. First, we pool all η^2-statistics together. Then we randomly split the pool into 'past' and 'future' groups of the same size as the data in Figure 2e and calculate the mean value in each group. We repeat this process 1,000,000 times and use this bootstrapped distribution to calculate a p-value for the original difference between means. This process is carried separately for 1st order sequence correlations and for ≥2nd order sequence correlations.

Peak location, onset location, and relative duration of sequence correlated activity (Figure 3d)—The data in Figure 3a is used to create the following 3 distributions:

Relative peak timing: The trial-averaged signals (rows in Figure 3a differ in ROIs and phrase type) are calculated after time-warping the signals to a fixed phrase duration, $T_{phrase} = 1$, whose onset is set to $T_{onset} = 0$. The timing of the signal peak, t_{peak}, is therefore already normalized because $t_{peak} = \frac{t_{peak} - T_{onset}}{T_{phrase}}$.

Relative onset timing: The signal in each trial contributing to Figure 3a is fitted with a hidden Markov model (as explained above in the "Seeking ROIs with sequence correlations" section). The onset time points of the signal state, t_{onset}, are normalized with respect to the phrase onset time, T_{onset}, and the phrase duration, T_{phrase}:

$$\hat{t}_{onset} = \frac{t_{onset} - T_{onset}}{T_{phrase}}$$

Relative signal duration: A threshold at 0.5 is used to identify segments of reliable state occupancy within the traces in Extended Data Fig. 7d. The resulting signal segments are in time normalized coordinates and represent the duration relative to the phrase duration.

Simulating point neuron fluorescence response to spike trains (Extended Data Fig. 7a)—To simulate the expected calcium indicator signal in response to a spike train, $sp(t)$, we use the empirical single-spike response:

$$K(t) = \begin{cases} \frac{1 - e^{-t/0.045}}{1 - e^{-1}} & 0 \le t \le 0.045 sec \\ e^{-(t - 0.045)/0.142} & t > 0.045 sec \end{cases}$$

Corresponding to a rise time constant of 45 msec and decay time constant of 142 msec (c.f. Supplementary table #3 in[25]). The above kernel is a low boundary on the rise time because it assumes 45msec for the full signal rise time and not just half-way. This is done to give a limit on what can be resolved.

For a point neuron we do not assume other dynamical processes that stem from morphology. The simulated signal is the convolution of the spike train with the kernel, K:

$$F(t) = \int_{-\infty}^{t} sp(\tau) K(t - \tau) d\tau$$

Contrasting influence of preceding and following phrases on neural activity (Extended Data Fig. 8g–i)—For neurons with significant sequence correlations (1-way ANOVA described above) we adopted a method agnostic to correlation order (1st or higher, as defined above) and direction (past or future). We used multi-way ANOVA to test the effect of the identity of the immediately preceding and immediately following phrase types on the neural signal ($s = \sum_{t \in p} (\Delta f / f_0)_{denoised}$). Using Tukey's post-hoc comparison and a threshold at p = 0.05 we compare the fractions of sequence-correlated ROIs influenced by past phrases, future phrases, or both. This comparison is also carried separately for ROIs active in complex transitions or outside of complex transitions (panels h,i).

Testing if sequence correlated neurons prefer one or more song contexts (Extended Data Fig. 9)—For neurons with significant sequence correlations (1-way ANOVA described above) we used Tukey's post-hoc analysis to determine if this sequence correlation results from a significant single preferred context or significant several preferred contexts. A neuron was declared 'single context preferring' if the mean signal in only that context was larger than all others (Tukey's $p<0.001$). A neuron was declared as having preference to more than a single past context if the mean signal following several contexts was larger than another context (Tukey's $p<0.001$). Since the post-hoc test uses a subset of

the songs it is weaker than the 1-way ANOVA and some neurons do not show a clear preference to one context or more but still have sequence correlation (gray in Extended Data Fig. 9f)

Maximum projection images for comparing context-dependent signals

Maximum fluorescence images: In songs that contain a fixed phrase sequence and a variable context element, such as a preceding phrase identity, maximum projection images are created, as above, but using only video frames from the target phrase (e.g. the pink phrase in Figure 2d). Then, the sets of maximum projection images in each context (e.g. identity of upstream phrase) are averaged, assigned orthogonal color maps (e.g. red and cyan in Extended Data Fig. 5) and overlaid. Consequentially, regions of the imaging plane that have no sequence preference will be closer to gray scale, while ROIs with sequence preference will be colored. In Extended Data Fig. 5 and Extended Data Fig. 9 we used a sigmoidal transform of the color saturation to amplify the contrast between color and gray scale without changing the sequence preference information. Additionally, to show that pixels in the ROI are biased towards the same context preference, the above context-averaged maximum projection images are subtracted and pseudo-colored (insets in Extended Data Fig. 5)

De-noised Maximum projection images (Figure 4a): The above maximum projection images show the fluorescence signal, including background levels that are typical to single-photon microscopy. To emphasize context-dependent ROIs we de-noised the fluorescence videos using the previously-published algorithm, CNMFE[49], and created maximum projection images, as above, from the background-subtracted videos. The preceding context preferring ROIs from this estimation algorithm (Figure 4a) completely overlap with the manually defined ROIs, used to extract signal rasters (Figure 4b). Extended Data Fig. 8j replicates Figure 4a without the de-noising algorithm and shows that the same ROIs report the same context dependence. SI video 8 shows all the de-noised video data, used to create Figure 4a.

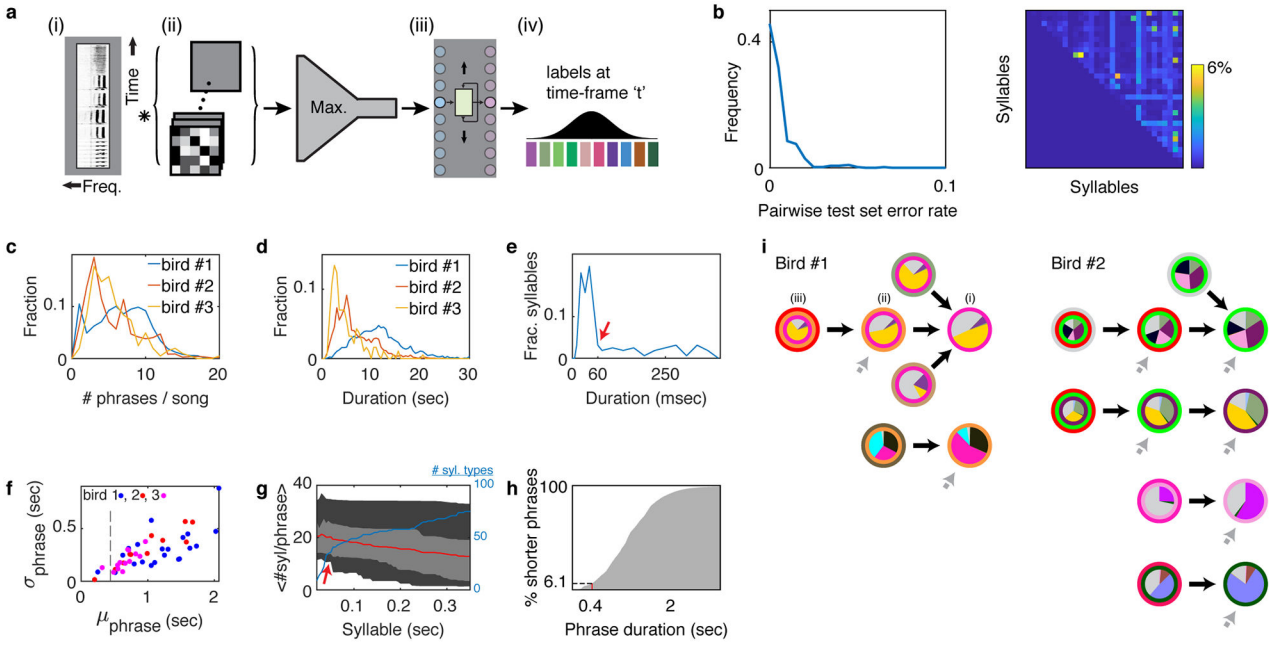
Label prediction from clustered network states—The signal integral during a target phrase, pink in Extended Data Fig. 10a, is used to create network states–vectors, composed of signals from 4 jointly-recorded ROIs. The averages of the vectors, belonging to the contexts defined by the 1st upstream (or downstream) phrase label define label-centroids. Then, labels of individual songs are assigned to the nearest neighboring centroid (Euclidean).

Bootstrapping mutual information in limited song numbers—The neurons in Extended Data Fig. 10a were recorded in 54 songs. This repetition number is too small for estimating the full distribution function of behavior and network activity states. To overcome this limitation, the mutual information between the network state and the identity of the 1st upstream (or downstream) phrase is estimated in a bootstrapping permutation process as follows:

- Sub-sampling 3 out of 4 ROIs in each permutation and converting their signal to binary values by thresholding the signal integral.

- Reducing the number of phrase labels by merging. Specifically, in Extended Data Fig. 10, the least common label in downstream states is randomly merged with one of the other labels. In the upstream labels, the least common label is merged after a random division of the other 4 labels, to form 2 groups of 2.
- The mutual information measures are then calculated for each one of the 48 possible state spaces and divided by the entropy of the behavior state – leading to the scatter Extended Data Fig. 10b. The margin of error is estimated from the standard deviation.
- The 0.95 quantile level of the null hypothesis is created by randomly shuffling each variable to create a 1000 surrogate data sets and repeating the measures.
- The shuffled set is used to create a sample distribution and calculate the significance of the differences in Extended Data Fig. 10b using a z-test with the sample mean and standard deviation.

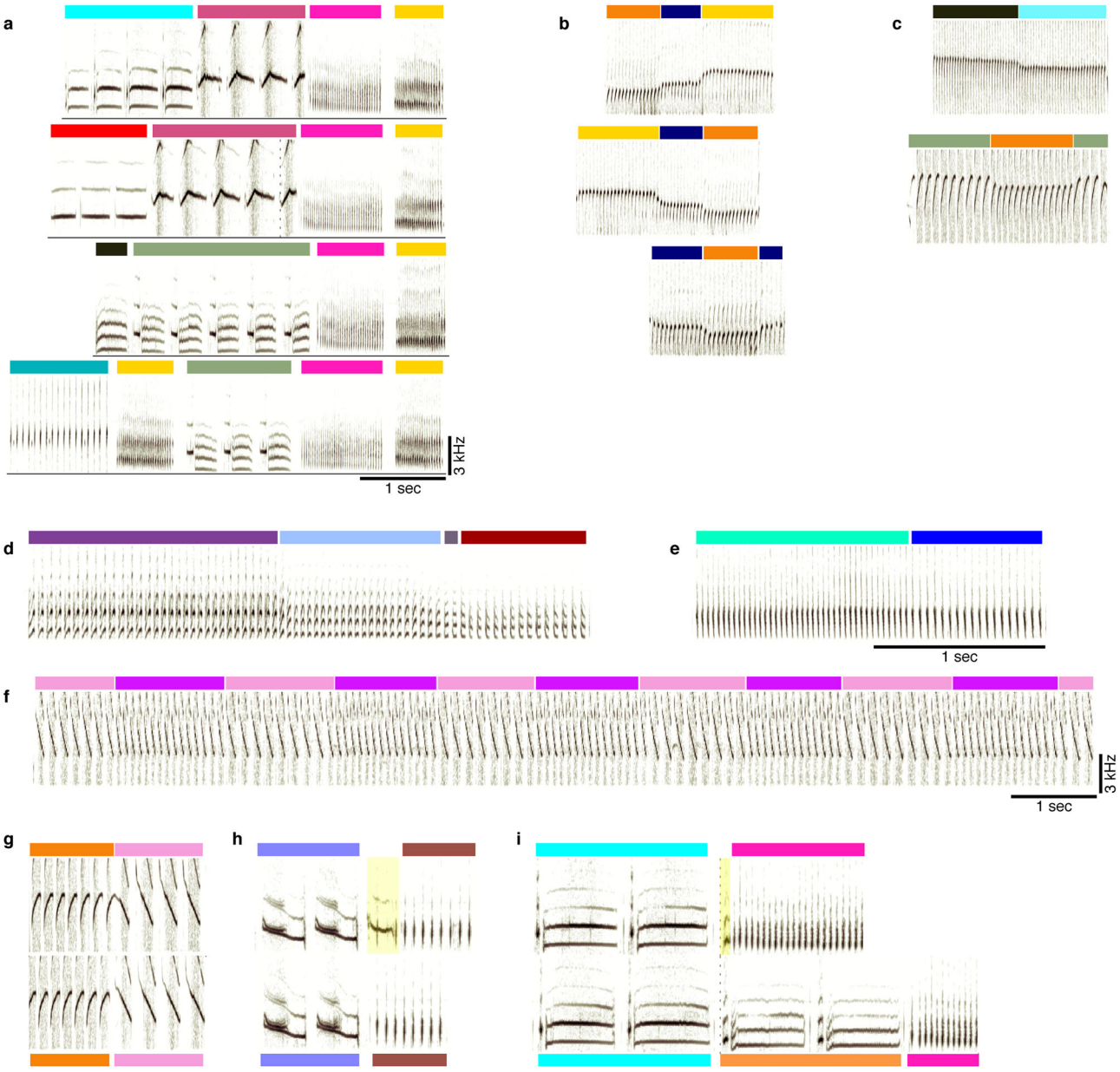
Extended Data



Extended Data Fig. 1 |. Canary song annotation and sequence statistics.

a. Architecture of syllable segmentation and annotation machine learning algorithm. (i) A spectrogram is fed to the algorithm as a 2D matrix in segments of 1 second. (ii). Convolutional and max-pooling layers learn local spectral and temporal filters. (iii). Bidirectional recurrent Long-Short-Term-Memory (LSTM) layer learns temporal sequencing features. (iv). Projection onto syllable classes assigns a probability for each 2.7 millisecond time bin and syllable. **b.** After manual proof reading (methods), a support vector machine (SVM) classifier was used to assess the pairwise confusion between all syllables classes of bird #1 (methods). The test set confusion matrix (right) and its histogram (left) show that in rare cases the error exceeded 1% and at most reached 6%. Since the higher values occurred only in phrases with 10s of syllables this metric guarantees that most of the syllables in every phrase cannot be confused as belonging to another syllable class. Accordingly, the possibility for making a mistake in identifying a phrase type is negligible. **c.** Histogram of the number of phrases per song for 3 birds used in this study. **d.** Histogram of song durations for 3 birds. **e.** Histogram of mean syllable durations, 85 syllable classes from 3 birds. Red arrow marks the duration, below which all trill types have more than 10 repetitions on average. **f.** Relation between phrase classes' duration mean (x-axis) and standard deviation (y-axis). Syllables classes (dots) of 3 birds are colored by the bird number. Dashed line marks 450 msec, an upper limit for the decay time constant of GCaMP6f. **g.** Range of mean number of syllables per phrase (y-axis) for all syllable types with mean duration shorter than the x-axis value. Red line is the median, light gray marks the 25%, 75% quantiles and dark gray mark the 5%, 95% quantile (blue line marks the # of syllable types contributing to these statistics). The red arrow matches the arrow in panel e. **h.** Cumulative histogram of trill phrase durations. **i.** All complex phrase transitions with ≥2nd order dependence on song history context (for birds #1, #2). For each phrase type that precedes a complex transition, the context dependence is visualized by a graph called a Probabilistic Suffix Tree (methods). Transition outcome probabilities are marked by pies at the center of each node. The song

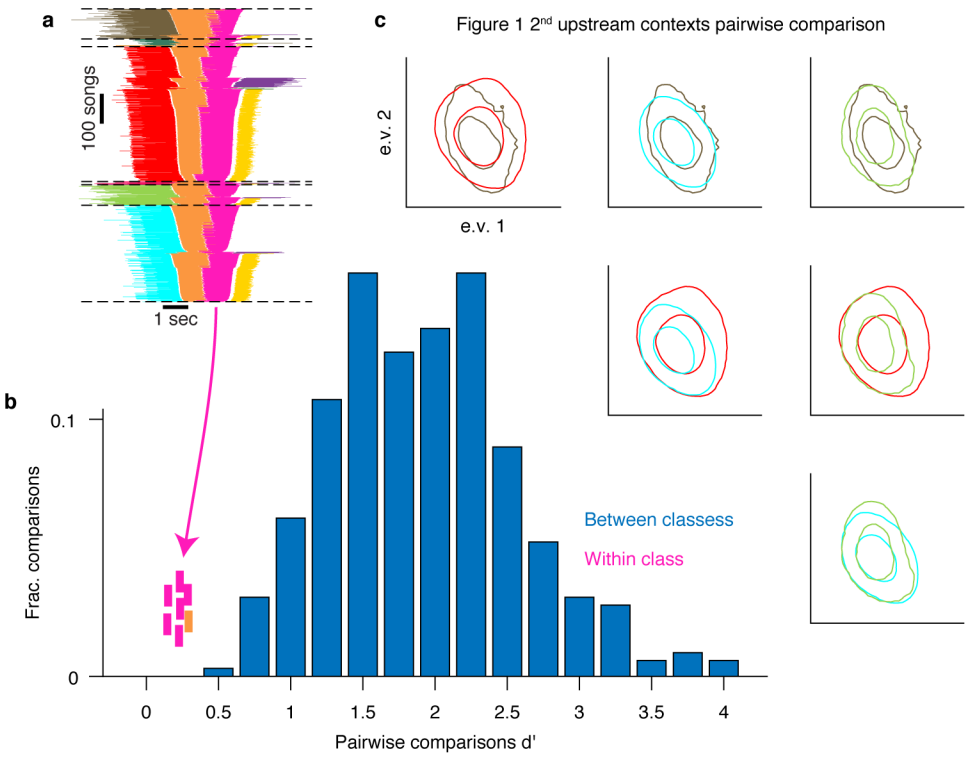
context—phrase sequence—that leads to the transition, is marked by concentric circles, the inner most being the phrase type preceding the transition. Nodes are connected to indicate the sequences in which they are added in the search for longer Markov chains that describe context dependence (e.g. i-iii for 1st to 3rd order Markov chains). Grey arrows indicate additional incoming links that are not shown for simplicity.



Extended Data Fig. 2 |. Examples of canary song phrase sequences, rare inter-phrase gaps, and aberrant syllables.

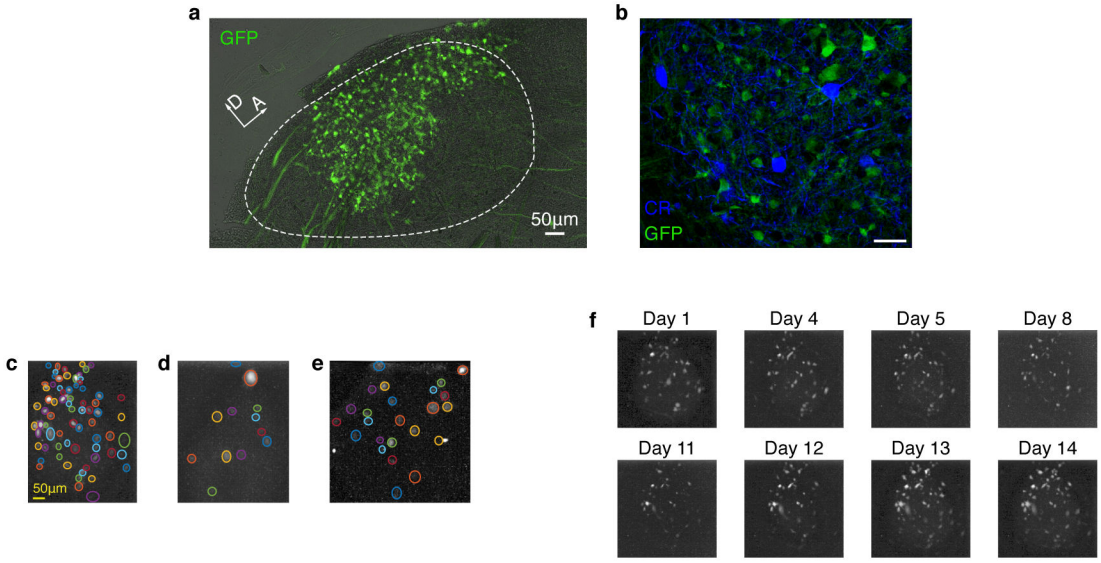
a. Additional spectrograms of phrase sequences (colors above the spectrograms indicate phrase identity), leading to a repeating pair of phrases (pink and yellow). **b.** Examples of flexible phrase sequencing comprised of pitch changes (from bird #3). **c.** Examples of phrase transitions with a pitch change from bird #2. **d-f.** Phrase sequences showing changes in spectral and temporal parameters. d, bird #1, changes from up sweep (purple) to down sweep (dark red) through intermediate phrases of intermediate acoustic structure. e, bird #1, a change in inter-syllable gaps. f, from bird #2, changes in pitch sweep rate. **g.** Top and bottom sonograms compare the same phrase transitions where the inter-phrase gap varies. **h,**

i. The top sonogram includes a rare vocalization in the beginning of the 2nd phrase (highlighted) that, in panel i, resemble the onset of an orange phrase type.



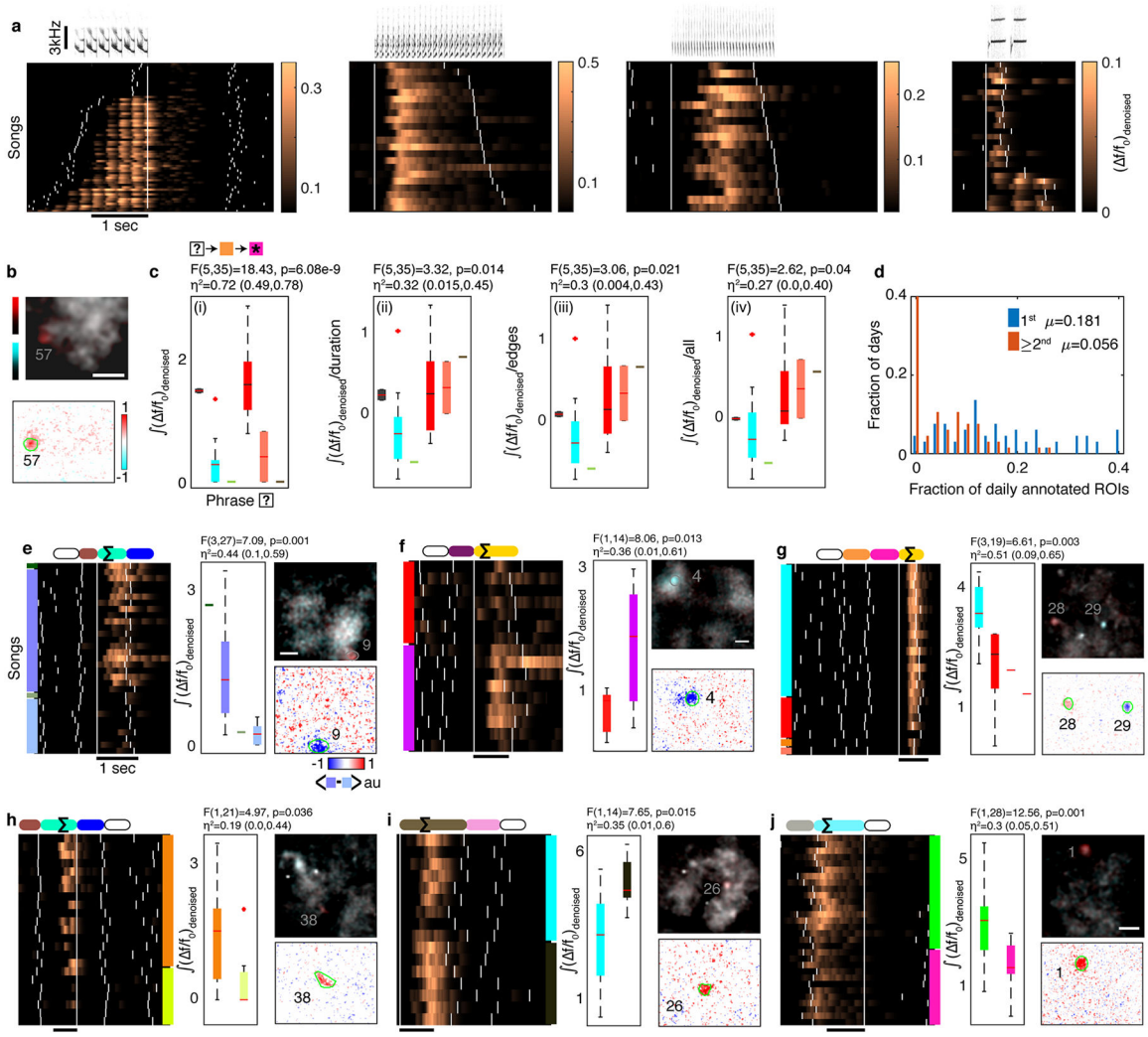
Extended Data Fig. 3 |. An example in which context-dependence of syllable acoustics prior to complex transitions is too small for clear distinction.

a. Repeats main figure 1b. A summary of all phrase sequences that contain a common transition reveals that the choice of what to sing after the pink phrase depends on the phrases that were produced earlier. Lines represent phrase identity and duration. Song sequences are stacked (vertical axis) sorted by the identity of the 1st phrase, the last phrase and then the center phrases' duration. **b.** The discriminability (d', x-axis) measures the acoustic distance between pairs of syllable classes in units of the within-class standard deviation (methods). Bars show the histogram across all pairs of syllables identified by human observers (methods) corresponding to about 99% or larger identification success (in Extended Data Fig. 1b). The pink ticks mark the d' values for 6 within-class comparison of the main 4 contexts in panel a. The orange tick marks the d' another context comparison in a different syllable that precedes a complex transition for this bird. **c.** The pairwise comparison of distributions matching the pink ticks in panel b. Each inset shows overlays of two distributions marked by contours at the 0.1 and 0.5 values of the peak and colored by the context in panel a. The distributions are projected onto the 2 leading principle components of the acoustic features (methods). While some of these distributions are statistically distinct they only allow for ~70% context identification success in the most distinct case.



Extended Data Fig. 4 |. Calcium indicator is expressed exclusively in HVC excitatory neurons and imaged in annotated regions of interest (ROIs)

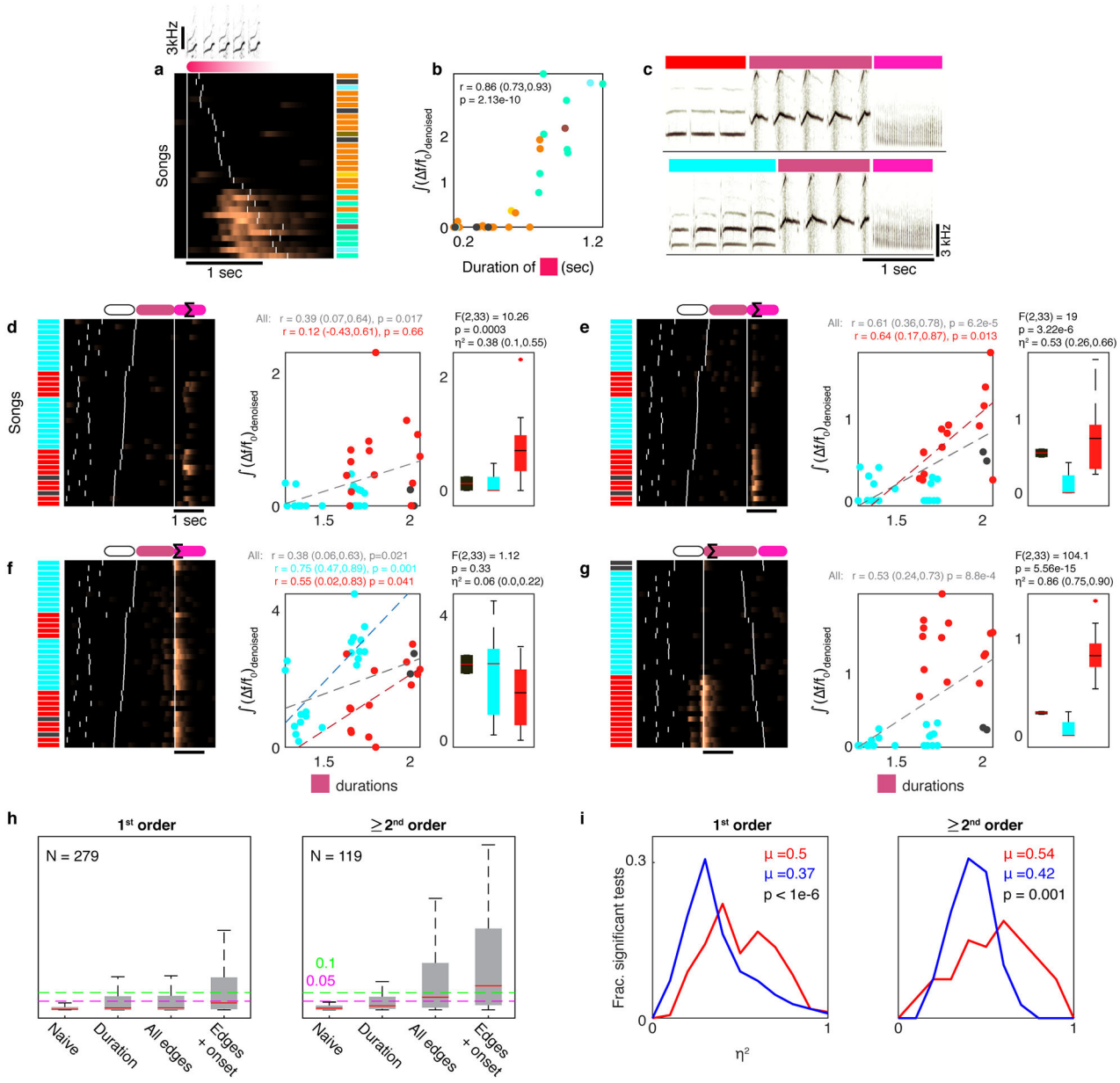
a. Sagittal slice of HVC showing GCaMP expressing projection neurons (Experiment repeated in 5 birds with similar results). **b.** We observed no overlap between transduced GCaMP6f-expressing neurons, and neurons stained for the inhibitory neurons markers calretinin, calbindin, and parvalbumin (CR stain shown, staining experiment repeated 6 times for each marker with similar results)**. c-e.** Example of daily ROI annotation in 3 birds. Colored circles mark different ROIs, manually annotated on maximum fluorescence projection images an exemplary day (see methods). Panel are for birds 1–3. **f.** Maximum fluorescence images (methods, from bird 1) revealing the fluorescence sources including sparsely active cells in the imaging window across multiple days.



Extended Data Fig. 5 |. Syllable and phrase-sequence-correlated ROIs from 3 birds.

a. Sonograms on top of rasters from 4 ROIs from 3 birds. White ticks indicate phrase onsets. The fluorescent calcium indicator is able to resolve individual long syllables. **b.** Top, average maximum fluorescence images during the pink phrase in Figure 2d, compare the two most common contexts in orthogonal colors (red and cyan). Scale bar is 50μm. Bottom, the difference of the overlaid images. ROI outlined in green. **c.** (i) 1-way ANOVA (F,p,η^2 and its 95% CI), tests the effect of contexts (x-axis, 2nd preceding phrase type in N=41 sequences) on the signal (y-axis. Lines, boxes, whiskers, and '+'s show the median, 1st and 3rd quartiles, full range, and outliers), during the target phrase (marked by ★) in Figure 2d. (ii-iv), ANOVA tests carried out using the residuals from the signal after removing the cumulative linear dependence on the duration of the target phrase, the relative timing of onset and offset edges of two fixed phrases, and the absolute onset time of the target phrase in each rendition. Colors correspond to phrases in Figure 2d. **d.** Histogram of fractions of daily annotated ROIs showing sequence correlation in all 3 birds. Each ROI can be counted only once per order. This estimate includes sparsely active ROIs. **e-j.** Activity during a target phrase (marked by ∑) is strongly related to non-adjacent phrase identities (empty ovals in color coded phrase

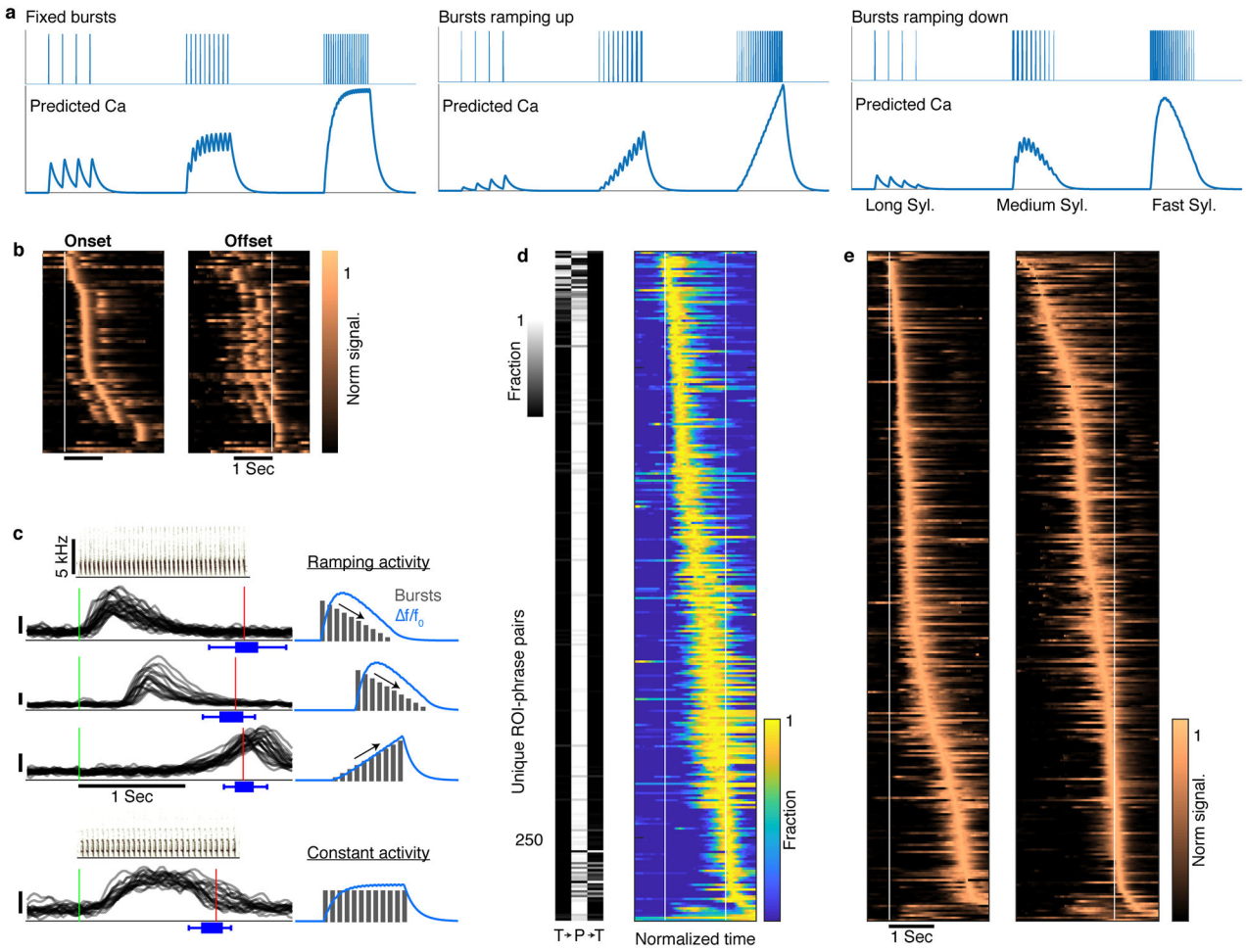
sequence). Songs are arranged by the phrase sequence context (left or right color patches for past and future phrase types). White ticks indicate phrase onsets. Box plots and contrast images as defined in panels b,c. N=31,16,23,23,16,30 songs contribute to panels e-j. **e,f.** Similar to main Figure 2d, $(\Delta f/f_0)_{denoised}$ from ROIs with 2nd order upstream sequence (color coded) from two more birds. **g.** 3rd order upstream relation. **h,i.** 2nd order downstream relations. **j.** 1st order downstream relation from another bird.



Extended Data Fig. 6 |. Phrases' durations and onset times also correlate to their sequence, but cannot fully account for HVC activity.

a. $(\Delta f/f_0)_{denoised}$ signal traces (ROI 18, bird 3) during one phrase type (red) arranged by its duration. Colored barcode annotates the final phrase in the sequence. **b.** The signal correlates to the red phrase's duration (r (95%CI), p: 2-sided Pearson's test for N=32 songs. Colors match barcode in panel a). **c.** Sonograms of two phrase sequences. **d-g.** ROI signals during N=36 sequences containing the last 2 phrases in panel c have various relations to the duration of the middle (purple) phrase (Scatter plots as in panel b. Dashed lines indicate significant correlations) and the identity of the 1st phrase (colors, 1-way ANOVA (F,p, η^2(95% CI)) tests the effect on the signal Σ. Whiskers, boxes, and lines show full range, 1st and 3rd quartiles, and medians). **d.** Signal correlation with phrase duration is completely entangled with the signal's sequence preference and does not apply in separate preceding

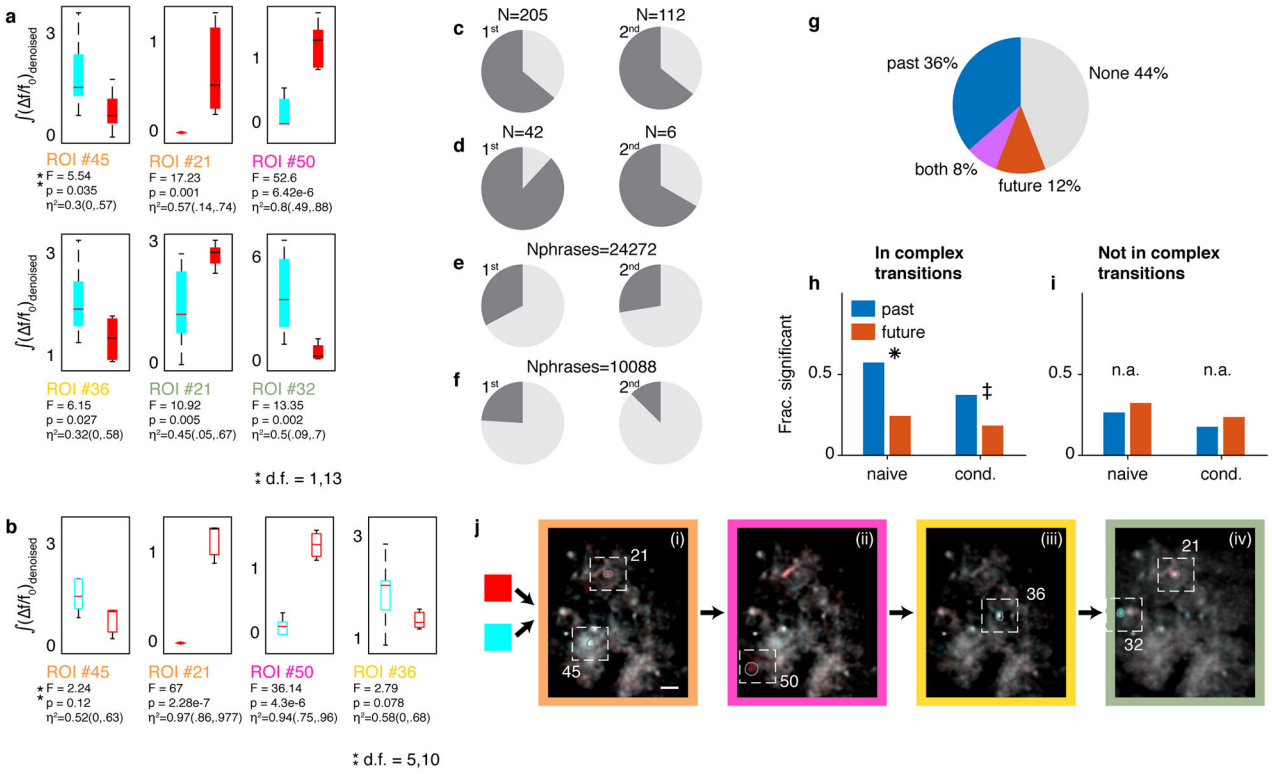
contexts (red, $p > 0.5$). **e.** Signal correlation with phrase duration is influenced by the signal's sequence preference but also exists in the preferred sequence context separately (red). **f.** Signal duration correlation is observed within each single preceding context separately, but the correlation reduces across all songs. **g.** Similar to panel a, but the signal is in the 2nd phrase, not the 3rd. **h.** Distributions of 1-way ANOVA p-values (y-axis, whiskers, boxes, and red lines show full range, 1st and third quartiles, and medians) relating phrase identity and signal for adjacent phrases (N=279 independent 1st order tests, left) and non-adjacent phrases (N=119 independent ≥2nd order tests, right). Tests are also done on residuals of signals, after discounting the following variables: variance explained by the target phrase duration, the timing of all phrase edges in the test sequence, and the time-in-song (x-axis, effects accumulated left to right by multivariate linear regression, see methods). Colored, dashed lines mark 0.05 and 0.1 p-values. **i.** Effect size (η^2 denotes frac. variance accounted for by the signals' context dependence) of past (red) and future (blue) 1-way ANOVA tests for 1st order (left, N=279 tests) and ≥2nd order (right, N=119) correlations. Difference of the mean value (μ) is tested using 1-sided bootstrap shuffles (p-values, methods).



Extended Data Fig. 7 |. Signal shape and onset time of sequence-correlated HVC neurons reflect within-phrase timing.

a. Simulation of calcium indicator (GCaMP6f) fluorescence corresponding to syllable-locked spike bursts in HVC projection neurons. Syllable-locked spike bursts are convolved with the indicator's kernel (methods) to estimate the expected signal when the number of spikes per burst is constant (left), ramps up (middle), or ramps down (right) linearly with the syllable number. The simulation assumes one burst per syllable in time spacing (x-axis) that matches long canary syllables (400–500msec), medium range syllables (100msec) and short syllables (50msec). **b.** Complementing Figure 3a, average context-sensitive activity in phrases with long syllables reveals syllable-locked peaks aligned to phrase onsets (left) or offsets (right, same row order as left) that change in magnitude across the phrase. **c.** Signal shape and onset timing has properties of within-phrase timing codes. Example raw $\Delta f/f_0$ signals (y-axis, 0.1 marked by vertical bar) of 4 ROIs aligned to onset of specific phrase types (green line, sonograms show the repeating syllables. Red lines and blue box plots show the median, range, and quartiles of the phrase offset timing). The signal shapes resemble the expected fluorescence of the calcium indicator elicited by syllable-locked ramping (sketches, top three) or constant activity. **d.** Left, barcode show the fraction of signal onsets found in the preceding transition, within the phrase, and in the following transition (T→P→T, methods). Rows correspond to the phrases in Figure 3a. Right, rows

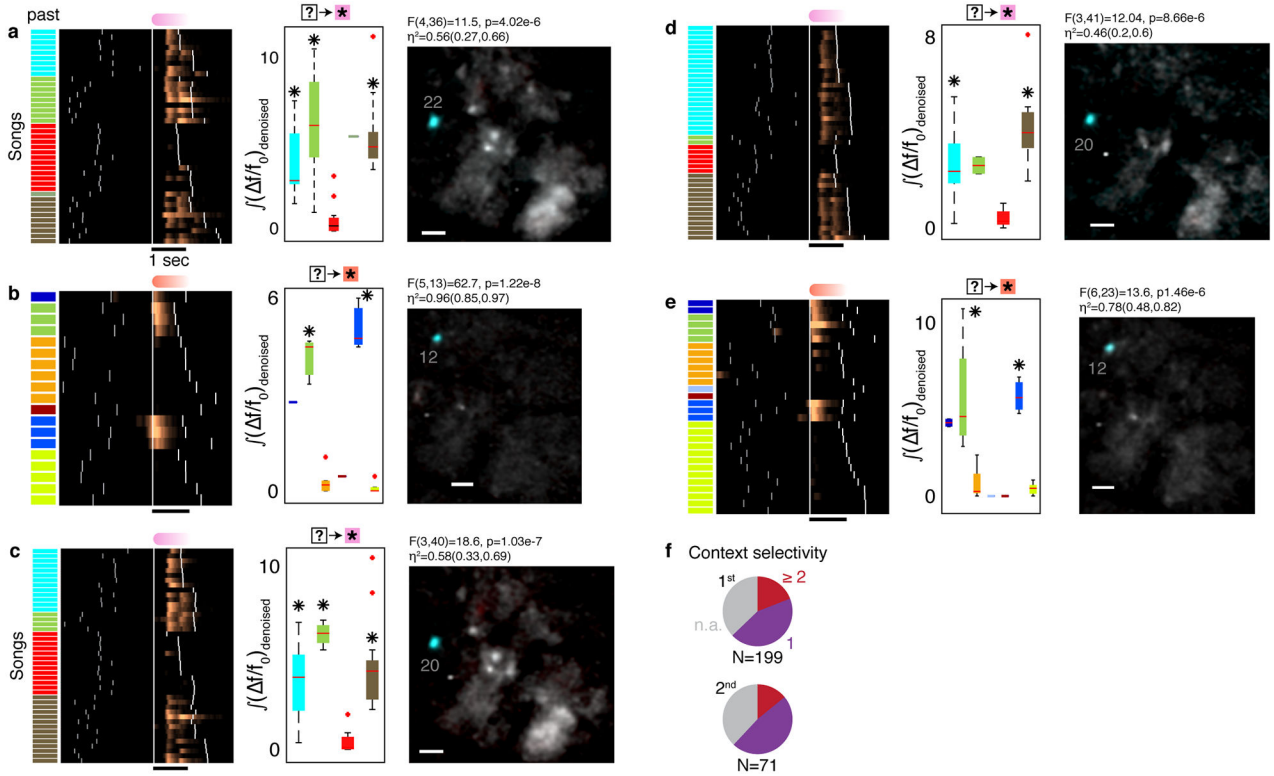
show the average signal state occupancy estimated from HMMs fitted to the single-trial data contributing to Figure 3a. The resulting traces are time-warped to fixed phrase edges (white lines). **e.** The single-trial data in Figure 3a is aligned to phrase onsets (left) and offsets (right) and averaged in real time. The resulting traces are ordered by peak location (separately in left and right rasters).



Extended Data Fig. 8 |. Context sensitive signals aggregate in complex transitions and preferentially encode past transitions.

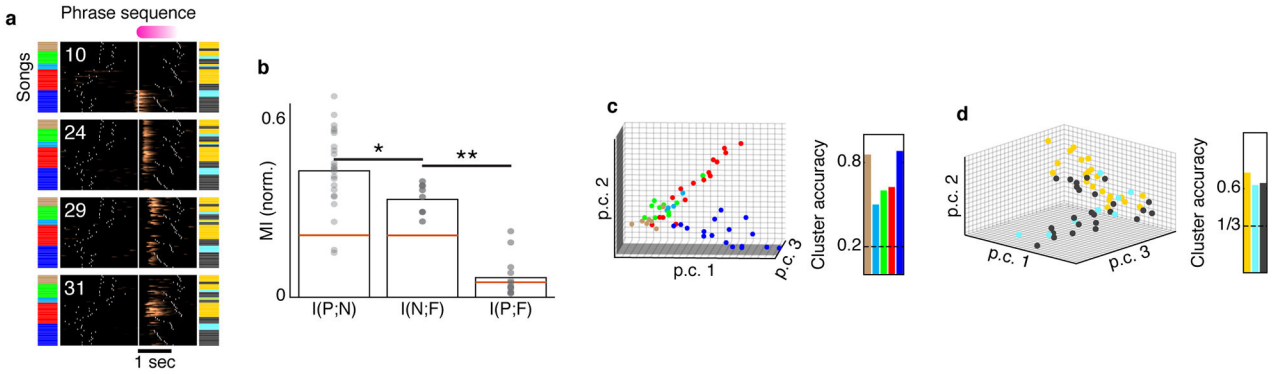
a. Distribution of signal integrals (y-axis, whiskers show full range, boxes show 1st and third quartiles, and lines show the medians) for ROIs in Figure 4a. (Text label is color coded by phrase type in sub-panels i-iv). F-numbers, p-values, and η^2 (95% CI) for 1-way ANOVA relating history (x-axis) and signal (y-axis) in N=15 song sequences. **b.** ROIs in panel (a) retain their song-context bias also for songs that happen to terminate at end of the third phrase rather than continue. Box plots repeat the ANOVA tests in panel (a) for N=16 songs in which the last phrase is replaced by end-of-song. **c-f.** Dark grey slices indicate the fraction of correlations occurring in complex behavioral transitions. **c,d.** the data in Figure 4c separated to the two birds. **e,f.** The fraction in panels c,d expected by the null hypothesis of correlations distributing by the frequency of each phrase type among $N_{phrases}$ phrases in the dataset. **g.** In sequence-correlated ROIs, multi-way ANOVA is used to separate the effect of the preceding and following phrase types on the signal (methods). Pie shows the percent of sequence-correlated ROIs significantly influenced by the past, future, or both phrase identities among N=336 significant ANOVA tests. **h.** Restricting analysis to complex transitions, more ROIs correlated to the preceding phrase type (blue) than to following (red). This is true in both Naive signal values (left, N=185 tests) and after removing dependencies on phrase durations and time-in-song (right, N=185). (one-sided binomial z-test: ✳: proportion difference 0.33 ± 0.09, Z=6.45, p = 5.5e-11, ‡: proportion difference 0.19 ± 0.09, Z=4.05,p = 2e-5). **i.** Restricting to phrase types not in complex transitions (N=136 ANOVA tests) reveals more ROIs correlated with the future phrase type but the difference is not significant (left,right n.a.: one-sided binomial z-test, p = 0.14,0.11). **j.** Figure 4a showed maximum projection images, calculated with de-noised videos (methods). The algorithm,

CNMF-E[49], involves estimating the source ROI shapes, de-convolving spike times as well as estimating the background noise. Here, recreating the maximum projection images with the original fluorescence videos shows the background as well but the preceding-context-sensitive neurons remain the same. Namely, the same ROI footprints annotated in panels i-iv show the color bias (cyan or red) that indicates coding of the past phrase with the same color.



Extended Data Fig. 9 |. ROIs reflecting several preceding song contexts.

a,b. ROIs active in multiple preceding contexts. $(\Delta f/f_0)_{denoised}$ traces are aligned to a specific phrase onset, arranged by identity of preceding phrase (color barcode). White ticks indicate phrase onsets. Box plot shows distributions of $(\Delta f/f_0)_{denoised}$ integrals (y-axis, summation in the phrase marked by ★) for various song contexts (x-axis). F-number, p-value, and effect size (η^2(95% CI)) show the significance of separation by song context (1-way ANOVA) and ✳ marks contexts that lead to larger mean activity compared to another context (Tukey's multiple comparisons, N=41 songs p=0.01,7.5e-6,5.6e-5 in a, N=19, p=8.8e-7,8.15e-8 in b). Average maximum projection images (methods) during the aligned phrase compare the song contexts that lead to significantly higher activity to the other contexts in orthogonal colors (cyan and red for high and low activity). Bar is 50μm. **c-e.** Neurons with similar context preference like the examples in panels a,b in adjacent days. (Tukey's multiple comparisons: N=44, p=0.001,4.08e-6,1.3e-6 in c. N=45, p=0.0016,2.85e-6 in d. N=30, p=0.0002,0.0001 in e). **f.** Fraction of ROIs with selectivity for one context (purple) or multiple contexts (red) identified using Tukey's post-hoc multiple comparisons (methods). Grey slices (n.a.) mark context-sensitive ROIs for which the post-hoc analysis did not isolate a specific context with larger mean signal. Top (bottom) pie shows selectivity for 1st (2nd) preceding phrases.



Extended Data Fig. 10 |. HVC neurons can be tuned to complementary preceding contexts.

a. Four jointly-recorded ROIs exhibit complementary context selectivity. Color bars indicate phrase identities preceding and following a fixed phrase (pink). For each ROI (rasters), $(\Delta f/f_0)_{denoised}$ traces are aligned to the onset of the pink phrase (x-axis) arranged by the identity of the preceding phrase, by the following phrase and finally by the duration of pink phrase. **b.** For the example in (a), normalized mutual information between the identity of past (P) and future (F) phrase types is significantly smaller than the information held by the network states about the past and future contexts (left bars. N is the 4-ROIs activity). Dots, bars, and red lines mark bootstrap assessment shuffles, their mean, and the 95% level of the mean in shuffled data (methods). *: difference is 0.09 ± 0.03, Z =4.3, p=7.3e-6, **: difference is 0.26 ± 0.02, Z=8.9, p< 1e-15, bootstrapped one-sided z-test. **c.** Signal integrals from the 4 ROIs in panel a are plotted for each song (dots, N=54 songs) on the 3 most informative principle components. Dots are colored by the identity of the preceding phrase. Clustering accuracy measures the 'leave-one-out' label prediction for each preceding phrase (true positive), calculated by assigning each dot to the nearest centroid (L_2). Dashed line marks chance level. **d.** Similar to panel c but for the 1st following phrase.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported by NIH grants R01NS089679, R01NS104925, R24NS098536 (T.J.G) and R24HL123828, U01TR001810 (D.K.)

The authors would like to thank Jeff Markowitz, Ian Davison, and Jeff Gavornik for useful discussions and comments to this manuscript.

The authors thank Nvidia Corporation for their technology grant (Y.C.)

References

1. Markowitz JE, Ivie E, Kligler L & Gardner TJ Long-range Order in Canary Song. PLOS Comput Biol 9, e1003052 (2013). [PubMed: 23658509]
2. Nottebohm F, Stokes TM & Leonard CM Central control of song in the canary, Serinus canarius. J. Comp. Neurol 165, 457–486 (1976). [PubMed: 1262540]
3. Hahnloser RHR, Kozhevnikov AA & Fee MS An ultra-sparse code underlies the generation of neural sequences in a songbird. Nature 419, 65–70 (2002). [PubMed: 12214232]

4. Long MA & Fee MS Using temperature to analyse temporal dynamics in the songbird motor pathway. Nature 456, 189–194 (2008). [PubMed: 19005546]
5. Rokni U, Richardson AG, Bizzi E & Seung HS Motor learning with unstable neural representations. Neuron 54, 653–666 (2007). [PubMed: 17521576]
6. Todorov E Optimality principles in sensorimotor control. Nat. Neurosci 7, 907–915 (2004). [PubMed: 15332089]
7. Wolpert DM Computational approaches to motor control. Trends Cogn. Sci 1, 209–216 (1997). [PubMed: 21223909]
8. Leonardo A Degenerate coding in neural systems. J. Comp. Physiol. A Neuroethol. Sens. Neural. Behav. Physiol 191, 995–1010 (2005). [PubMed: 16252121]
9. Jin DZ & Kozhevnikov AA A Compact Statistical Model of the Song Syntax in Bengalese Finch. PLOS Comput. Biol 7, e1001108 (2011). [PubMed: 21445230]
10. Ohbayashi M, Ohki K & Miyashita Y Conversion of Working Memory to Motor Sequence in the Monkey Premotor Cortex. Science 301, 233–236 (2003). [PubMed: 12855814]
11. Goldman-Rakic PS Cellular basis of working memory. Neuron 14, 477–485 (1995). [PubMed: 7695894]
12. Svoboda K & Li N Neural mechanisms of movement planning: motor cortex and beyond. Curr. Opin. Neurobiol 49, 33–41 (2018). [PubMed: 29172091]
13. Thompson JA, Costabile JD & Felsen G Mesencephalic representations of recent experience influence decision making. eLife 5,.
14. Pastalkova E, Itskov V, Amarasingham A & Buzsáki G Internally generated cell assembly sequences in the rat hippocampus. Science 321, 1322–1327 (2008). [PubMed: 18772431]
15. Churchland MM, Afshar A & Shenoy KV A central source of movement variability. Neuron 52, 1085–1096 (2006). [PubMed: 17178410]
16. Mushiake H, Saito N, Sakamoto K, Itoyama Y & Tanji J Activity in the Lateral Prefrontal Cortex Reflects Multiple Steps of Future Events in Action Plans. Neuron 50, 631–641 (2006). [PubMed: 16701212]
17. Shima K & Tanji J Neuronal Activity in the Supplementary and Presupplementary Motor Areas for Temporal Organization of Multiple Movements. J. Neurophysiol 84, 2148–2160 (2000). [PubMed: 11024102]
18. Fujimoto H, Hasegawa T & Watanabe D Neural Coding of Syntactic Structure in Learned Vocalizations in the Songbird. J. Neurosci 31, 10023–10033 (2011). [PubMed: 21734294]
19. Hamaguchi K, Tanaka M & Mooney R A Distributed Recurrent Network Contributes to Temporally Precise Vocalizations. Neuron 91, 680–693 (2016). [PubMed: 27397518]
20. Ashmore RC, Wild JM & Schmidt MF Brainstem and Forebrain Contributions to the Generation of Learned Motor Behaviors for Song. J. Neurosci 25, 8543–8554 (2005). [PubMed: 16162936]
21. Alonso RG, Trevisan MA, Amador A, Goller F & Mindlin GB A circular model for song motor control in Serinus canaria. Front. Comput. Neurosci 9, (2015).
22. Goldberg JH & Fee MS Singing-related neural activity distinguishes four classes of putative striatal neurons in the songbird basal ganglia. J. Neurophysiol 103, 2002–2014 (2010). [PubMed: 20107125]
23. Jin DZ Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus HVC. Phys. Rev. E Stat. Nonlin. Soft Matter Phys 80, 051902 (2009). [PubMed: 20365001]
24. Hinton G et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Process. Mag 29, 82–97 (2012).
25. Chen T-W et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature 499, 295–300 (2013). [PubMed: 23868258]
26. Bouchard KE & Brainard MS Auditory-induced neural dynamics in sensory-motor circuitry predict learned temporal and sequential statistics of birdsong. Proc. Natl. Acad. Sci 113, 9641–9646 (2016). [PubMed: 27506786]

27. Wittenbach JD, Bouchard KE, Brainard MS & Jin DZ An Adapting Auditory-motor Feedback Loop Can Contribute to Generating Vocal Repetition. PLoS Comput. Biol 11, e1004471 (2015). [PubMed: 26448054]
28. Dave AS, Yu AC & Margoliash D Behavioral State Modulation of Auditory Activity in a Vocal Motor System. Science 282, 2250–2254 (1998). [PubMed: 9856946]
29. Cardin JA & Schmidt MF Noradrenergic Inputs Mediate State Dependence of Auditory Responses in the Avian Song System. J. Neurosci 24, 7745–7753 (2004). [PubMed: 15342742]
30. Glaze CM & Troyer TW Development of temporal structure in zebra finch song. J. Neurophysiol 109, 1025–1035 (2013). [PubMed: 23175805]
31. Castelino CB & Schmidt MF What birdsong can teach us about the central noradrenergic system. J. Chem. Neuroanat 39, 96–111 (2010). [PubMed: 19686836]
32. Prather JF, Peters S, Nowicki S & Mooney R Precise auditory–vocal mirroring in neurons for learned vocal communication. Nature 451, 305–310 (2008). [PubMed: 18202651]
33. Okubo TS, Mackevicius EL, Payne HL, Lynch GF & Fee MS Growth and splitting of neural sequences in songbird vocal development. Nature 528, 352–357 (2015). [PubMed: 26618871]
34. Zucker RS & Regehr WG Short-term synaptic plasticity. Annu. Rev. Physiol 64, 355–405 (2002). [PubMed: 11826273]
35. Iacobucci GJ & Popescu GK NMDA receptors: linking physiological output to biophysical operation. Nat. Rev. Neurosci 18, 236–249 (2017). [PubMed: 28303017]
36. Nagel K, Kim G, McLendon H & Doupe A A bird brain's view of auditory processing and perception. Hear. Res 273, 123–133 (2011). [PubMed: 20851756]
37. Fiete IR, Senn W, Wang CZH & Hahnloser RHR Spike-Time-Dependent Plasticity and Heterosynaptic Competition Organize Networks to Produce Long Scale-Free Sequences of Neural Activity. Neuron 65, 563–576 (2010). [PubMed: 20188660]
38. Abeles M Corticonics: Neural Circuits of the Cerebral Cortex. (Cambridge University Press, 1991).
39. Cannon J, Kopell N, Gardner T & Markowitz J Neural Sequence Generation Using Spatiotemporal Patterns of Inhibition. PLOS Comput. Biol 11, e1004581 (2015). [PubMed: 26536029]
40. Hamaguchi K & Mooney R Recurrent interactions between the input and output of a songbird cortico-basal ganglia pathway are implicated in vocal sequence variability. J. Neurosci. Off. J. Soc. Neurosci 32, 11671–11687 (2012).
41. Graves A, Mohamed A & Hinton G Speech recognition with deep recurrent neural networks. in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing 6645–6649 (2013). doi:10.1109/ICASSP.2013.6638947.
42. Yamashita Y & Tani J Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: A Humanoid Robot Experiment. PLoS Comput. Biol 4, (2008).
43. Santoro A et al. Relational recurrent neural networks in Advances in Neural Information Processing Systems 31 (eds. Bengio S et al.) 7310–7321 (Curran Associates, Inc., 2018).
44. Chorowski JK, Bahdanau D, Serdyuk D, Cho K & Bengio Y Attention-Based Models for Speech Recognition in Advances in Neural Information Processing Systems 28 (eds. Cortes C, Lawrence ND, Lee DD, Sugiyama M & Garnett R) 577–585 (Curran Associates, Inc., 2015).
45. Stokes TM, Leonard CM & Nottebohm F The telencephalon, diencephalon, and mesencephalon of the canary, Serinus canaria, in stereotaxic coordinates. J. Comp. Neurol 156, 337–374 (1974). [PubMed: 4609173]
46. Liberti Iii WA et al. Unstable neurons underlie a stable learned behavior. Nat. Neurosci 19, 1665–1671 (2016). [PubMed: 27723744]
47. Wild JM, Williams MN, Howie GJ & Mooney R Calcium-binding proteins define interneurons in HVC of the zebra finch (Taeniopygia guttata). J. Comp. Neurol 483, 76–90 (2005). [PubMed: 15672397]
48. Wohlgemuth MJ, Sober SJ & Brainard MS Linked Control of Syllable Sequence and Phonology in Birdsong. J. Neurosci 30, 12936–12949 (2010). [PubMed: 20881112]
49. Zhou P et al. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. eLife 7, e28728 (2018). [PubMed: 29469809]

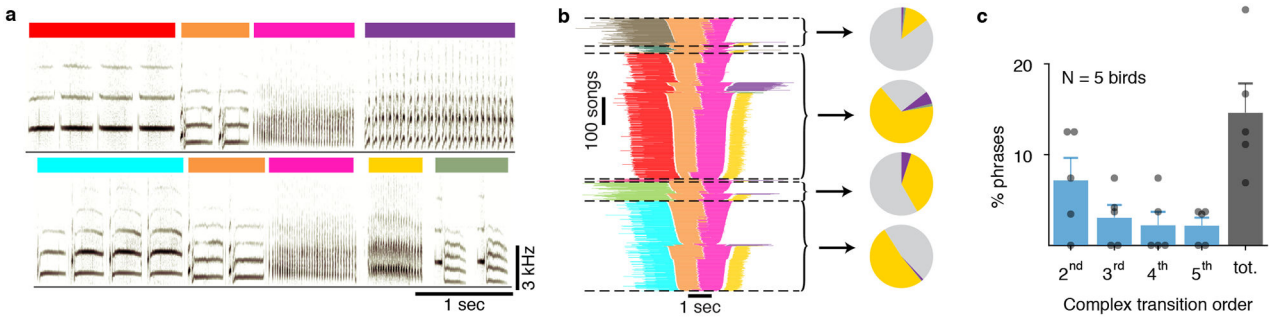


Figure 1 |. Long range syntax rules in canary song.

a. Two example spectrograms of canary song. Colored bars indicate different phrases assembled from basic elements called syllables. Both examples contain a common phrase transition (orange to pink) but differ in the preceding and following phrases. **b.** A summary of all phrase sequences containing this common transition reveals that the choice of what to sing after the pink phrase depends on the phrases that were produced earlier. Lines represent phrase identity and duration. Song sequences are stacked (vertical axis) sorted by the identity of the 1st phrase, the last phrase and then the center phrases' duration. Pie charts show the frequency of phrases that follow the pink phrase, calculated in the subset of songs that share a preceding sequence context (separated by dashed lines). In the pie chart, grey represents the song end, and other colors represent a phrase pictured in the first panel. The pink phrase precedes a 3rd order 'complex transition'; the likelihood that a particular phrase will follow it is dependent on transitions three phrases in the past. **c.** Percent of phrases that precede complex transitions of different orders in N=5 birds (dots). Bars and error bars show mean and SE.

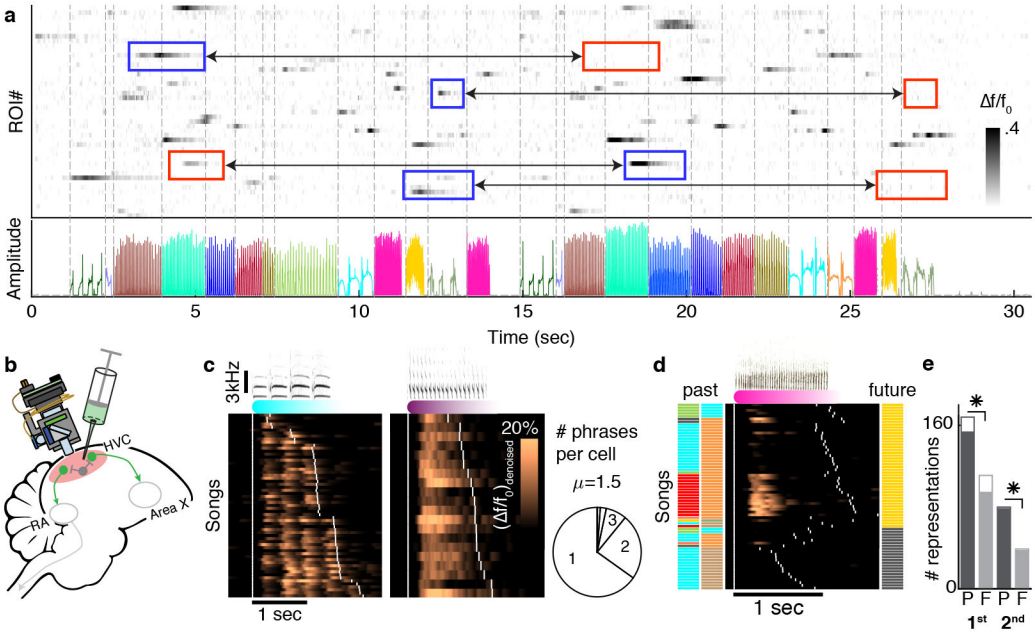


Figure 2 |. HVC projection neuron activity reflects long-range phrase sequence information.

a. Fluorescence ($\Delta f/f_0$) of multiple ROIs during a singing bout reveals sparse, phrase-type-specific activity. Phrase types are color coded in the audio amplitude trace, and dashed lines mark phrase onsets. Context-dependent ROIs show larger phrase-specific signal in one context (blue frames) than another (connected red frames). **b.** Experimental paradigm. Miniature microscopes were used to image GCaMP6f-expressing neurons in HVC, transduced via lentivirus injection. **c.** Most ROIs are phrase-type-specific. Neural activity is aligned to the onset of phrases. These phrases have long (left) and short (right) syllables and traces are sorted (y-axis) by the phrase duration. White ticks indicate phrase onsets. Pie shows fractions of ROIs that are active during just one, two or three phrase types (methods). **d.** Phrase-type-specific ROI activity that is strongly related to 2nd upstream phrase identity. Neural activity is aligned to the onset of the current phrase. Songs are arranged by the ending phrase identity (right, color patches), then by the phrase sequence context (left, color patches), and then by duration of the pink phrase. White ticks indicate phrase onsets. **e.** Cells reveal more information about past events than future events. 307 different ROIs had 398 significant correlations with adjacent (1st order, 2 left bars) and non-adjacent (≥2nd order, 2 right bars) phrases. The correlations are separated by phrases that precede (P) or follow (F) the phrase, during which the signal is integrated. Empty bars mark transition-locked representations (methods, Extended Data Fig. 7d). 2-sided binomial z-test evaluate significant differences (✳: proportion differences 0.2 ± 0.08, 0.34 ± 0.11, Z=4.82,5.31, p=1.39e-6, 1.065e-7 for 1st and ≥2nd order).

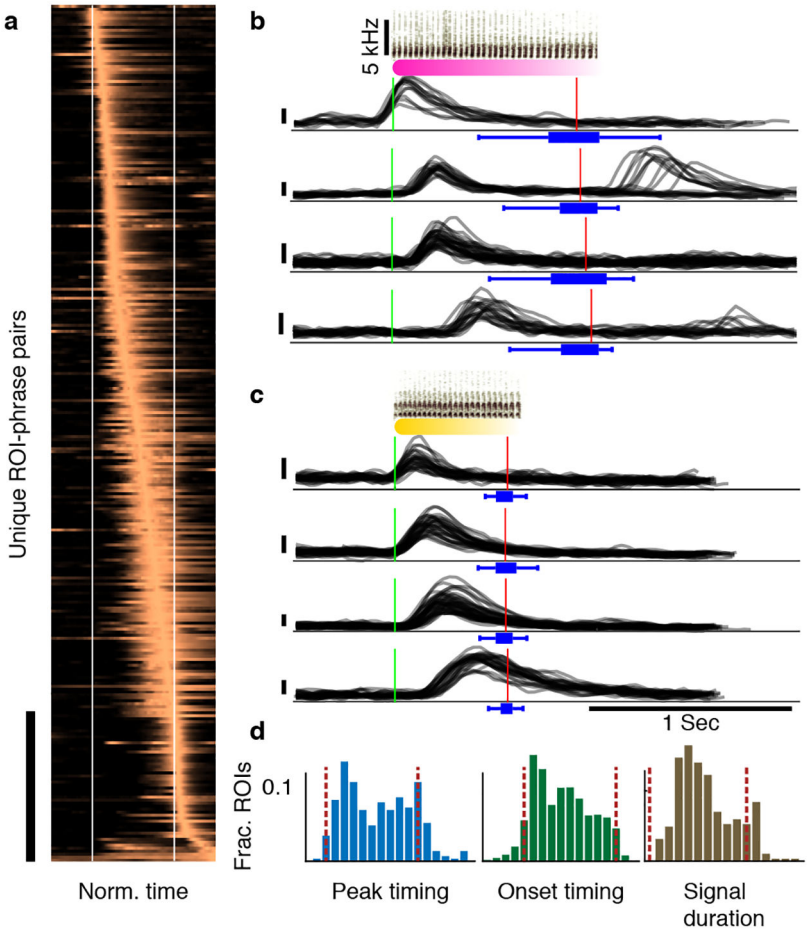


Figure 3|. Sequence-correlated HVC neurons reflect within-phrase timing.

a. Activity of context-sensitive ROIs (y-axis, bar marks 50 rows) is time-warped to fixed phrase edges (x-axis, white lines) and averaged across repetitions of short-syllable phrases. Traces are ordered by their peak timing to reveal the span of the phrase time frame. **b,c.** Example raw $\Delta f/f_0$ traces (y-axis, vertical bars equal 0.1) of 8 ROIs during phrase types that precede (b) and follow (c) the complex transition in Figure 1. Traces are aligned to phrase onsets (green line, sonograms show the syllables) and panels show ROIs with various onset timing across the phrase. Red lines and blue box plots show the median, range, and quartiles of the phrase offset timing (top to bottom: N = 70,23,55,39,40,38,50,31 phrases summarized by the box plots). **d.** Histograms showing the distribution of peak timing (left), onset timing (middle) and signal durations (right) of the activity in panel a relative to the phrase edges (dashed lines).

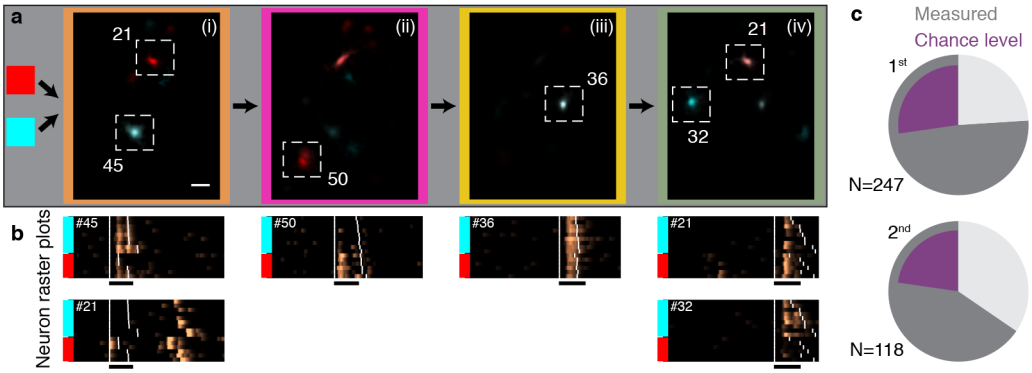


Figure 4 |. Sequence-correlated HVC neurons reflect preceding context up to four phrases apart and show enhanced activity during context-dependent transitions.

a. A sequence of four phrases (i-iv, color coded) is preceded by two upstream phrase types (red or cyan). Average maximum projection denoised images (methods) are calculated in each sequence context during each phrase in the sequence (i-iv) and overlaid in complementary colors (red, cyan) to reveal context-preferring neurons. Scale bar is 50 μm. **b.** $(\Delta f/f_0)_{denoised}$ rasters for the ROIs in panel (a). Songs are ordered by the preceding phrase type (colored bars). Extended Data Fig. 8a shows the statistical significance of song context relations. **c.** Fraction of sequence-correlated ROIs found in complex transitions. Pie charts separate 1st order and higher order (≥2nd) sequence correlations. Dark grey summarizes the total fraction for two birds. Purple shows fractions expected from sequence correlates uniformly-distributed in all phrase types.