Article

Hidden neural states underlie canary song syntax

https://doi.org/10.1038/s41586-020-2397-3

Received: 24 February 2019

Accepted: 26 March 2020

Published online: 17 June 2020

Check for updates

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[4,5,6], Darrell N. Kotton[4,5,6] & Timothy J. Gardner[1,7]✉

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

Canary songs, like many flexible behaviours, contain complex transitions—points at which the next action depends on memory for choices made several steps in the past. Songs are composed of syllables produced in trilled repetitions known as phrases (Fig. 1a) that are about 1 s long and are sung in sequences, typically 20–40 s long. The order of phrases in a song exhibits 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 (Fig. 1b, c).

In premotor brain regions, neural activity that supports long-range complex transitions will reflect context information as redundant representations of ongoing behaviour[5–8]. Such representations, referred to here as 'hidden neural states', have been predicted in models of memory-guided behaviour 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 that are 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 PNs fire in a way that depends on neighbouring 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 100 ms). By contrast, correlations in human behaviour can extend for tens of seconds and beyond, and are consistent with long-range syntax rules. At present it is not known whether redundant premotor representations in songbirds can support working memory for syntax control over timescales longer than 100 ms.

To further dissect the mechanisms of working memory for song we used custom head-mounted miniature microscopes to record HVC PNs during song production in freely moving canaries (*Serinus canaria*) (Fig. 2b). Although PNs can be divided into 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 undescribed pattern of neural dynamics that can support structured, context-dependent song transitions and validate predictions of long-range syntax generated by hidden neural states[9,23] in a complex vocal learner.

Complex transitions in a subset of phrases

Inspired by technological advances in human speech recognition[24], we developed a song segmentation and annotation algorithm that automated working with large data sets (more than 5,000 songs; Extended Data Fig. 1a, Methods). The birds' repertoire included 24–37 different syllables with typical durations of 10–350 ms. The average number of syllable repeats per phrase type ranged from 1 to 38, with extreme cases of individual phrases exceeding 10 s and 120 syllables (Extended Data

[1]Department of Biology, Boston University, Boston, MA, USA. [2]Boston University Center for Systems Neuroscience, Boston, MA, USA. [3]Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA, USA. [4]Center for Regenerative Medicine of Boston University and Boston Medical Center, Boston, MA, USA. [5]The Pulmonary Center, Boston University School of Medicine, Boston, MA, USA. [6]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [7]Phil and Penny Knight Campus for Accelerating Scientific Impact, University of Oregon, Eugene, OR, USA. ✉e-mail: ycohen1@mgh.harvard.edu; timg@uoregon.edu

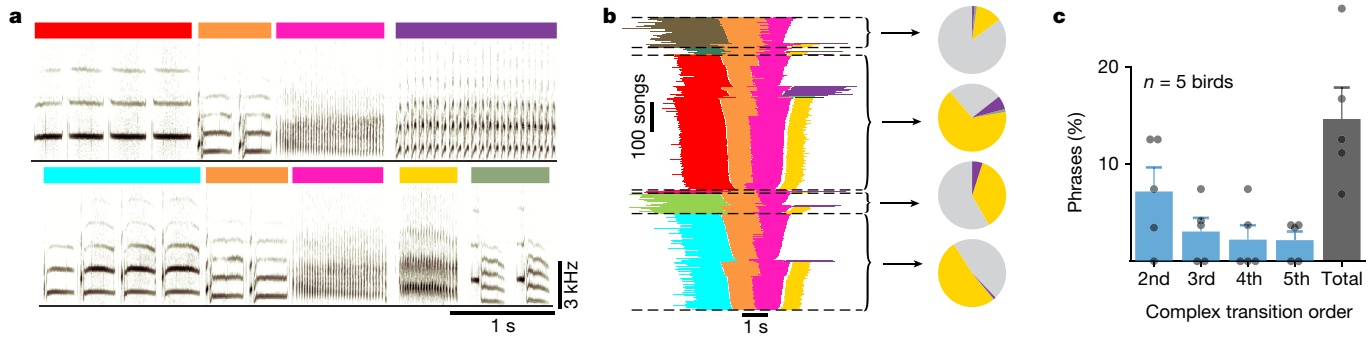


Fig. 1 | Long-range syntax rules in canary song. a, Two example spectrograms of canary song. Coloured bars indicate different phrases assembled from basic elements called syllables. Both examples contain a common phrase transition (orange to pink) but differ in the preceding and following phrases. **b**, 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) and ordered by the identity of the first phrase, the identity of the last phrase, and then the duration of the centre phrases. 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); grey represents the song end and other colours represent a phrase pictured in the left panel. The pink phrase precedes a third-order 'complex transition'; the likelihood that a particular phrase will follow it is dependent on transitions three phrases in the past. **c**, Percentage (mean + s.e.m.) of phrases that precede complex transitions of different orders in $n = 5$ birds (dots).

Fig. 1c–g). Transitions between phrases could be completely deterministic, where one phrase type always followed another, or flexible, where multiple phrase types could follow a given phrase (Fig. 1a, b). In very rare cases, transitions contained an aberrant syllable that could not be stably classified (Extended Data Fig. 2g–i), and all data were visually proofed. (Extended Data Figures 1b and 2 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—behavioural 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 three preceding phrases (Fig. 1b). Such song context dependence is captured by a third-order Markov chain. Extended Data Figure 1i shows the long-range context-dependent transitions for two birds.

HVC neurons encode long-range syntax

To characterize the neural activity that supports complex transitions, we imaged neurons that expressed the genetically encoded calcium indicator GCaMP6f in freely behaving adult male canaries ($n = 3$, age at least one year, recording in left hemisphere HVC[2]). The indicator is selectively expressed in PNs and neural activity to be recorded via fluorescence dynamics extracted from annotated regions of interest (ROIs; Extended Data Fig. 4, Methods). In our data set, 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[25] (400 ms; Extended Data Fig. 1h). As in finches, HVC PN activity in canaries was sparse in time[3,18]. Out of $n = 2{,}010$ daily annotated ROIs (mean ± s.d. of 35 ± 15 ROIs per animal per day), about 90% were selectively active in just one or two phrase types (Fig. 2a, c, Extended Data Fig. 5a). This, combined with the long phrase duration (Extended Data Fig. 1f, h), allowed us to examine the song-context dependence of neural activity using GCaMP6f. In our analysis, we treat recordings from different days separately. This approach overestimates the number of independent neurons we imaged but avoids analysis biases and stability concerns. Under the more conservative assumption that sources persist across days, in Supplementary Note 1 we still estimate 1,057 independent sources in our data set.

When we examined the patterns of phrase-locked activity, we identified signals that changed depending on song context. For example, some ROIs showed weak or no activity in one song context but demonstrated strong activity in another song context (Fig. 2a). Notably, this context-dependent activity was strongly influenced by the identity of non-adjacent phrases. For example, Fig. 2d shows the denoised fluorescence signal raster from a ROI, locked to the phrase type marked in pink, which displays a marked variation in activity ($(\Delta f/f_0)_{\text{denoised}}$, Methods) depending on the second phrase in the sequence's past—a second-order correlation. This sequence preference was quantified by integrating the ROI-averaged signal (Extended Data Fig. 5b, c; one-way ANOVA, $F_{5,35} = 18.3$, $P < 1 \times 10^{-8}$; one-way ANOVA evaluates the null hypothesis that there is no activity variation with phrase identity for all sequence-correlated ROIs in this manuscript). We found ROIs with signals that related to the identities of past and future non-adjacent phrases in all three birds (Extended Data Fig. 5). Across all birds, 21.2% of the daily annotated ROIs showed sequence correlations that extended beyond the current active syllable. In 18.1% there were first-order correlations, where activity during one phrase depends on the identity of an adjacent phrase, and in 5.6% there were second-order or greater 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] (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 effects of these duration and timing variables on the neural signals. We found that 32.8% (39/119 from 3 birds) of second-order or greater relations and 52.7% (147/279 from 3 birds) of first-order relations remained significant (Extended Data Figs. 5c, 6h).

The sequence-correlated ROIs tend to reflect past events more often than future events. Out of $n = 398$ significant correlations between neural activity and phrase sequence, 62.3% reflected preceding phrase identities (binomial z-test rejects the hypothesis of 50%, $z = 6.94$, $P < 1 \times 10^{-11}$). This bias was also found separately in first- or higher-order correlations (Fig. 2e, 60.2% and 67.2%, respectively; both percentages are significantly larger than 50%; binomial z-test, $z = 4.82$, 5.31 and $P < 1 \times 10^{-6}$, $P < 1 \times 10^{-6}$, respectively, and oppose the bias of 44.6% and 43.1% first- and second-order correlations expected to reflect past events from behaviour statistics alone; $P < 1 \times 10^{-7}$, binomial tests) and persisted

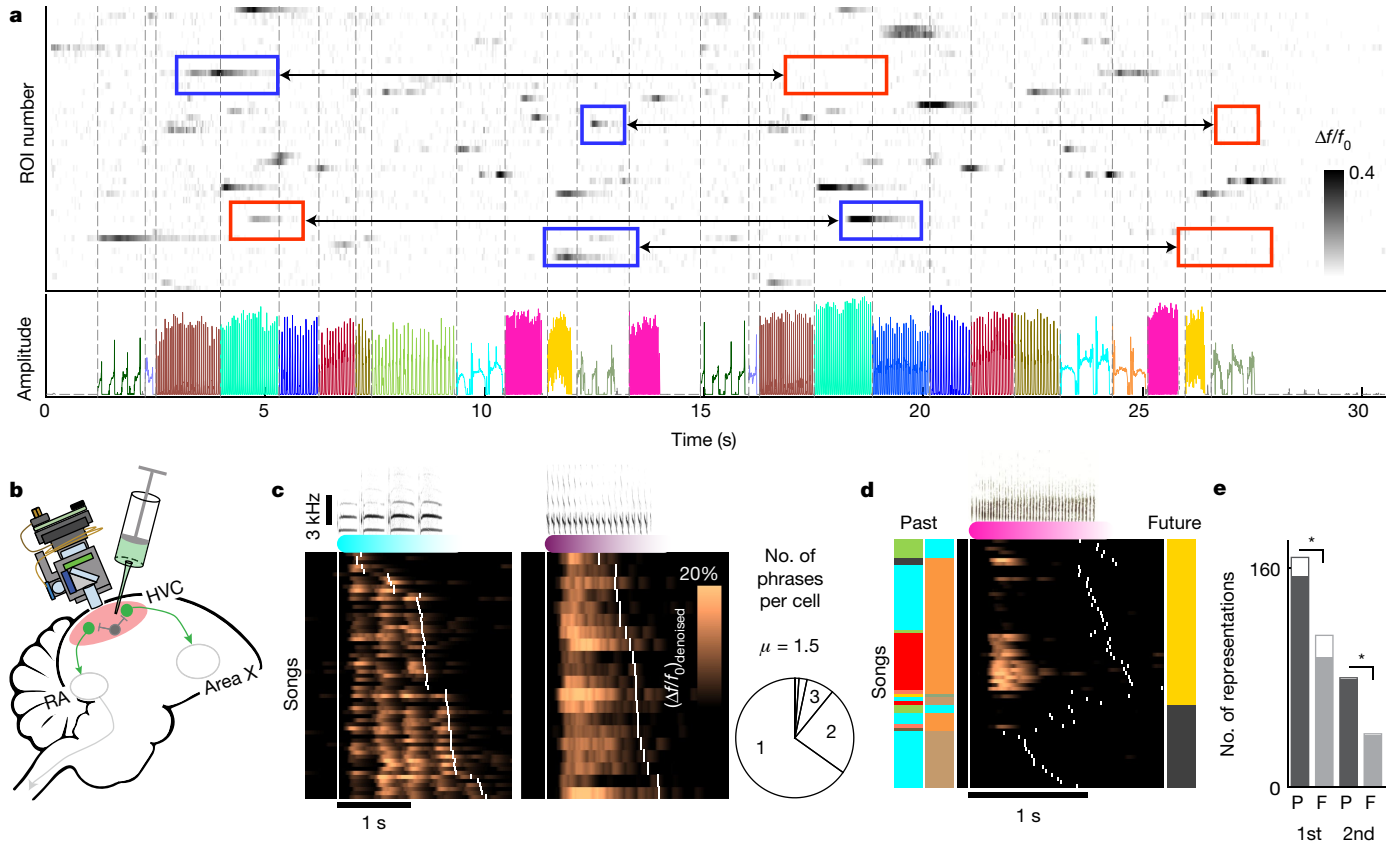


Fig. 2 | HVC PN activity reflects long-range phrase sequence information. **a**, Fluorescence ($\Delta f/f_0$) of multiple ROIs during a singing bout reveals sparse, phrase-type-specific activity. Phrase types are colour coded in the audio amplitude trace (bottom), and dashed lines mark phrase onsets. Context-dependent ROIs show larger phrase-specific signal in one context (blue frames) than another (connected red frames). **b**, Experimental setup. 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 phrase duration. White ticks indicate phrase onsets. Pie chart shows fractions of ROIs that are active during just one, two or three phrase types (see Methods). **d**, Phrase-type-specific ROI activity that is strongly related to second upstream phrase identity. Neural activity is aligned to the onset of the current phrase. Songs are arranged by the ending phrase identity (right, colour patches), then by the phrase sequence context (left, colour 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. Three-hundred and seven different ROIs had 398 significant correlations with adjacent (first order, two left bars) and non-adjacent (second or greater order, two 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 (see Methods, Extended Data Fig. 7d). Two-sided binomial *z*-test to evaluate significant differences (*proportion differences 0.2 ± 0.08 and 0.34 ± 0.11, $z = 4.82$ and 5.31, $P = 1.39 \times 10^{-6}$ and 1.065×10^{-7} for first and second or greater order, respectively).

when we considered ROIs that overlapped in footprint and sequence correlation across days as the same source (Supplementary Note 1). 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, $P < 1 \times 10^{-6}$ and $P = 0.001$ for first- and higher-order correlations, respectively).

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

HVC PNs also encode within-phrase timing

HVC PNs 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 finches, the same pattern was described for some cells as well as ramping syllable-locked spike bursts that increased or decreased in spike number over the course of a phrase[18]. In our data set, a small subset of ROIs was consistent with fixed syllable-locked neural activity (Fig. 2c, Extended Data Fig. 7a, c). More commonly, the activity was restricted to a brief period of time within a phrase, as in Fig. 2d, not time-locked to each syllable within the phrase. When we examined all sequence-correlated ROIs, we found that 91% were active for time-intervals shorter than the phrase, with peak timing and onset timing that can be found at all times in the phrase (Fig. 3, Extended Data Fig. 7b, c, e; also showing that 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 carry 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 of this process. For example, during a fixed sequence of four phrases, we found ROIs that carried forward information about the identity of the first phrase during each subsequent phrase (Fig. 4a, b, Extended Data Fig. 8a; one-way ANOVA showing significant modulation of neural activity with the identity of the past phrase). In this example, the ROIs that reflect long-range

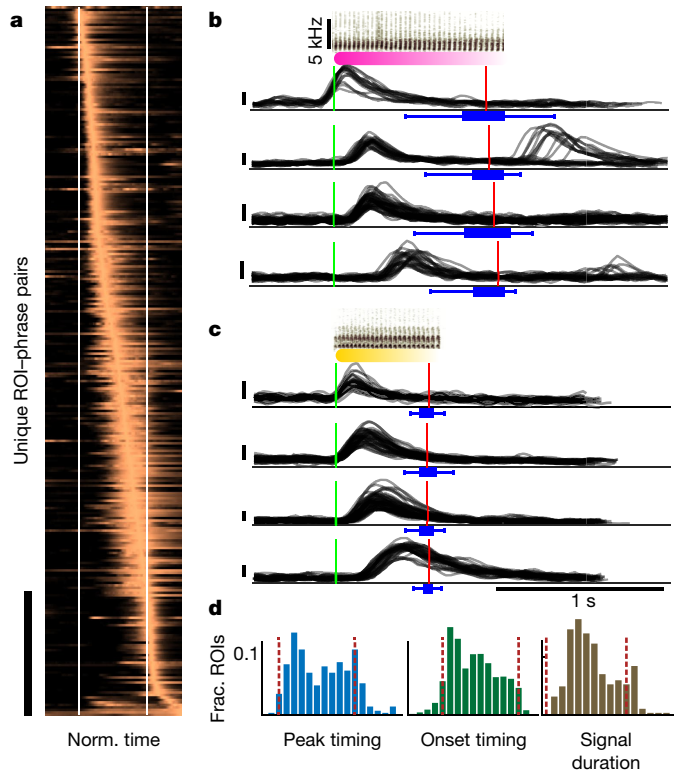


Fig. 3 | Sequence-correlated HVC neurons reflect within-phrase timing. **a**, Activity of context-sensitive ROIs (*y* axis, bar marks 50 rows) is time-warped to fixed phrase edges (*x* axis, white lines) and averaged across repetitions of short-syllable phrases. Traces are ordered by their peak timing 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 eight ROIs during phrase types that precede (**b**) or follow (**c**) the complex transition in Fig. 1. Traces are aligned to phrase onsets (green line; sonograms show 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: $n = 70, 23, 55, 39, 40, 38, 50$ and 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 **a** relative to the phrase edges (dashed lines).

information continue to do so even if the final phrase in the sequence is replaced by the end of the song, suggesting that their activity reflects prior song context rather than some upcoming future syllable choice (Extended Data Fig. 8b; one-way ANOVA, $F_{5,10} = 36.14$ and 2.79, $P < 5 \times 10^{-6}$ and $P < 0.08$ for ROIs 50 and 36, respectively, when replacing the last phrase with the end of song). This example suggests that a chain of neurons that reflect hidden states or information about past choices could provide the necessary working memory to implement long-range transition rules.

HVC neurons active in complex transitions

The phrases in Fig. 4 are phrase types that lead to complex transitions or directly follow them (in Fig. 1). If HVC neurons with context-selective activity are driving long-range syntax rules, then they should represent song context information predominately around complex behaviour 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 preceded or followed complex transitions (Fig. 4c; 76% (65%) for first (second or greater) order). Both percentages are larger than the 27% (22%) expected from uniform distribution of sequence-correlations in all phrases (binomial test, $P < 1 \times 10^{-10}$, Extended Data Fig. 8c–f) and persist if we consider ROIs that overlap in footprint and sequence correlation across days as the same source (Supplementary Note 1). When we separated the influence of past context and future action on the neural activity we found that, in complex transitions, ROIs predominately represented the identity of the preceding phrase (Extended Data Fig. 8g, h; multi-way ANOVA 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: $Z = 6.45$, $P < 1 \times 10^{-10}$). This bias does not occur outside 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.

Ensemble activity predicts complex behaviour

Of the ROIs with first-order and second- or greater-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 that respond in multiple contexts can complement each other to provide additional information about song history (Figs. 2d, 4 (ROIs 21, 45, 50)). Extended Data Figure 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 behaviour 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; see 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, $z = 8.95$, $P < 1 \times 10^{-15}$). 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 $z = 5.7$, $P < 1 \times 10^{-8}$, 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, $z = 4.32$, $P < 1 \times 10^{-5}$), suggesting that information is lost during the complex transition.

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

Discussion

Motor skills with long-range sequence dependencies are common in complex behaviours, with speech the richest example. In general, the neural mechanisms that underlie long-range motor sequence dependencies are unknown. Here we show that context-sensitive activity in HVC PNs can support the long-range order in canary song sequences[1]. Specifically, we find PNs the activity of which is contingent on phrases up to four steps in the past and PNs that predict phrases two steps into the future. Cells with this higher-order behaviour tend to be active during complex behavioural transitions—times at which the song behaviour requires high-level information about the sequence context. A key next step will be to further subdivide the activity reported here, in order to determine which PN classes in HVC carry the long-range information.

The HVC activity described here resembles the many-to-one relation between neural activity and behaviour states[9,23,27,33] proposed in some models to relay information across time. In this respect, our findings expand on a previous study in Bengalese finches[18] that identified HVC

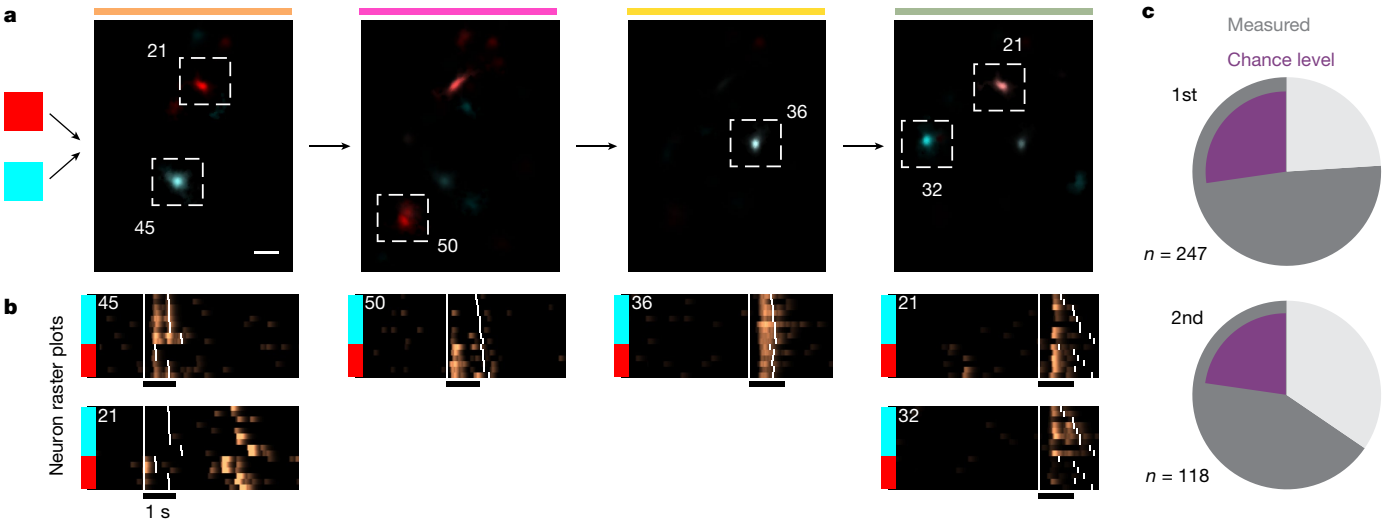


Fig. 4 | Sequence-correlated HVC neurons reflect preceding context up to four phrases apart and show enhanced activity during context-dependent transitions. a, A sequence of four phrases (left to right, colour coded) is preceded by two upstream phrase types (red or cyan). Average maximum projection denoised images (see Methods) are calculated in each sequence context during each phrase in the sequence and overlaid in complementary colours (red, cyan) to reveal context-preferring neurons. Scale bar, 50 µm. **b**, Raster plots of $(\Delta f/f_0)_{\text{denoised}}$ for the ROIs in **a**. Songs are ordered by the preceding phrase type (coloured bars). Extended Data Figure 8a shows the statistical significance of song context relations. Scale bars, 1 s. **c**, Fraction of sequence-correlated ROIs found in complex transitions. Pie charts separate first-order and higher-order sequence correlations. Dark grey summarizes the total fraction for two birds. Purple shows fractions expected from sequence correlates uniformly distributed in all phrase types.

PNs the activity of which 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 100 ms.

In the canary HVC neurons observed here, the time frame extends over multiple phrases and several seconds. This longer time frame rules out explanations based on short-term biophysical processes such as 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 50–250 ms time frame. Unlike the 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 that was predicted by a prominent statistical model of birdsong[9].

There are clues that HVC does not contain all of the information required to select a phrase transition—as more neurons correlate to the sequence's past than to its future, it is possible that sequence information in HVC is lost, perhaps owing 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 that are commonly added in sequence generating models[37–39], or may enter downstream, as well-documented noise in the basal ganglia outputs[40] also converges on pre-motor cortical areas downstream of HVC and may affect 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 behaviours in mammals. Finally, we note that recent marked progress in speech recognition algorithms has used recurrent neural networks with several architectures designed to capture sequence dependencies with hidden states. Examples include long short-term memory (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 to inform new models of many-to-one, history-dependent mappings between brain state and behaviour[23].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2397-3.

1. Markowitz, J. E., Ivie, E., Kligler, L. & Gardner, T. J. Long-range order in canary song. *PLOS Comput. Biol.* **9**, e1003052 (2013).
2. Nottebohm, F., Stokes, T. M. & Leonard, C. M. Central control of song in the canary, *Serinus canarius*. *J. Comp. Neurol.* **165**, 457–486 (1976).
3. Hahnloser, R. H. R., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419**, 65–70 (2002).
4. Long, M. A. & Fee, M. S. Using temperature to analyse temporal dynamics in the songbird motor pathway. *Nature* **456**, 189–194 (2008).
5. Rokni, U., Richardson, A. G., Bizzi, E. & Seung, H. S. Motor learning with unstable neural representations. *Neuron* **54**, 653–666 (2007).
6. Todorov, E. Optimality principles in sensorimotor control. *Nat. Neurosci.* **7**, 907–915 (2004).
7. Wolpert, D. M. Computational approaches to motor control. *Trends Cogn. Sci.* **1**, 209–216 (1997).
8. Leonardo, A. Degenerate coding in neural systems. *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.* **191**, 995–1010 (2005).
9. Jin, D. Z. & Kozhevnikov, A. A. A compact statistical model of the song syntax in Bengalese finch. *PLOS Comput. Biol.* **7**, e1001108 (2011).
10. Ohbayashi, M., Ohki, K. & Miyashita, Y. Conversion of working memory to motor sequence in the monkey premotor cortex. *Science* **301**, 233–236 (2003).
11. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
12. Svoboda, K. & Li, N. Neural mechanisms of movement planning: motor cortex and beyond. *Curr. Opin. Neurobiol.* **49**, 33–41 (2018).
13. Thompson, J. A., Costabile, J. D. & Felsen, G. Mesencephalic representations of recent experience influence decision making. *eLife* **5**, e16572 (2016).
14. Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsáki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321**, 1322–1327 (2008).
15. Churchland, M. M., Afshar, A. & Shenoy, K. V. A central source of movement variability. *Neuron* **52**, 1085–1096 (2006).
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).

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).
18. Fujimoto, H., Hasegawa, T. & Watanabe, D. Neural coding of syntactic structure in learned vocalizations in the songbird. *J. Neurosci.* **31**, 10023–10033 (2011).
19. Hamaguchi, K., Tanaka, M. & Mooney, R. A distributed recurrent network contributes to temporally precise vocalizations. *Neuron* **91**, 680–693 (2016).
20. Ashmore, R. C., Wild, J. M. & Schmidt, M. F. Brainstem and forebrain contributions to the generation of learned motor behaviors for song. *J. Neurosci.* **25**, 8543–8554 (2005).
21. Alonso, R. G., Trevisan, M. A., Amador, A., Goller, F. & Mindlin, G. B. A circular model for song motor control in *Serinus canaria*. *Front. Comput. Neurosci.* **9**, 41 (2015).
22. Goldberg, J. H. & Fee, M. S. Singing-related neural activity distinguishes four classes of putative striatal neurons in the songbird basal ganglia. *J. Neurophysiol.* **103**, 2002–2014 (2010).
23. Jin, D. Z. Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus HVC. *Phys. Rev. E* **80**, 051902 (2009).
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).
26. Bouchard, K. E. & Brainard, M. S. Auditory-induced neural dynamics in sensory–motor circuitry predict learned temporal and sequential statistics of birdsong. *Proc. Natl Acad. Sci. USA* **113**, 9641–9646 (2016).
27. Wittenbach, J. D., Bouchard, K. E., Brainard, M. S. & Jin, D. Z. An adapting auditory-motor feedback loop can contribute to generating vocal repetition. *PLOS Comput. Biol.* **11**, e1004471 (2015).
28. Dave, A. S., Yu, A. C. & Margoliash, D. Behavioral state modulation of auditory activity in a vocal motor system. *Science* **282**, 2250–2254 (1998).
29. Cardin, J. A. & Schmidt, M. F. Noradrenergic inputs mediate state dependence of auditory responses in the avian song system. *J. Neurosci.* **24**, 7745–7753 (2004).
30. Glaze, C. M. & Troyer, T. W. Development of temporal structure in zebra finch song. *J. Neurophysiol.* **109**, 1025–1035 (2013).
31. Castelino, C. B. & Schmidt, M. F. What birdsong can teach us about the central noradrenergic system. *J. Chem. Neuroanat.* **39**, 96–111 (2010).
32. Prather, J. F., Peters, S., Nowicki, S. & Mooney, R. Precise auditory–vocal mirroring in neurons for learned vocal communication. *Nature* **451**, 305–310 (2008).
33. Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F. & Fee, M. S. Growth and splitting of neural sequences in songbird vocal development. *Nature* **528**, 352–357 (2015).
34. Zucker, R. S. & Regehr, W. G. Short-term synaptic plasticity. *Annu. Rev. Physiol.* **64**, 355–405 (2002).
35. Iacobucci, G. J. & Popescu, G. K. NMDA receptors: linking physiological output to biophysical operation. *Nat. Rev. Neurosci.* **18**, 236–249 (2017).
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).
37. Fiete, I. R., Senn, W., Wang, C. Z. H. & Hahnloser, R. H. R. Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity. *Neuron* **65**, 563–576 (2010).
38. Abeles, M. *Corticonics: Neural Circuits of the Cerebral Cortex* (Cambridge Univ. 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).
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.* **32**, 11671–11687 (2012).
41. Graves, A., Mohamed, A. & Hinton, G. Speech recognition with deep recurrent neural networks. *2013 IEEE Intl Conf. Acoustics, Speech and Signal Processing* 6645–6649 (2013).
42. Yamashita, Y. & Tani, J. Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLOS Comput. Biol.* **4**, e1000220 (2008).
43. Santoro, A. et al. in *Advances in Neural Information Processing Systems 31* (eds Bengio, S. et al.) 7310–7321 (Curran Associates, 2018).
44. Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. & Bengio, Y. in *Advances in Neural Information Processing Systems 28* (eds Cortes, C. et al.) 577–585 (Curran Associates, 2015).

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).

Birds

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

Anaesthesia and analgesia. Before the birds were anaesthetized, they were injected with meloxicam (intramuscular, 0.5 mg/kg) and deprived of food and water for a minimum of 30 min. Birds were anaesthetized 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) was 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°, −2.5 mm ML, 0.12 mm AP, 0.15–0.7 mm D; nucleus RA: +80°, −2.5 mm ML, −1.2 mm AP, 1.9–3 mm D; area X: +20°, −1.27, −1.3 mm ML, 5.65, 5.8 mm AP, 2.65–2.95 mm 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 increases with depth.

HVC demarcation and head anchoring. To target HVC, 50–100 nl of the retrograde lipophilic tracer DiI (5 mg/ml solution in dimethylformamide, DMF) was injected into the left area X. The outer bone leaflet was removed above area X using a dental drill. The inner bone leaflet was thinned and removed using an ophthalmic scalpel, exposing a hole of ~300 μm diameter. The left area X was injected using a Drummond Nanoject II (Drummond pipette, 23 nl/s, pulses of 2.3 nl). In the same surgery, a head anchoring structure was created by curing dental acrylic (Flow-It ALC, Pentron) above the exposed skull and through ~100-μm 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 an area of the dura about 1.5–2 mm in diameter. The DiI demarcation of HVC was used to select an area for imaging. The lentivirus was injected in 3 or 4 locations, at least 0.2 mm apart, at a range of depths between 0.5 and 0.15 mm. In total 800–1,000 nl was injected into the left HVC. After the injection, the dura was removed and the parahippocampus segment above the imaging site was removed using 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 to cover the surgery area. The birds were allowed to recover for 1–2 weeks.

Hardware

To image calcium activity in HVC PNs during singing, we used custom, lightweight (~1.8 g), commutable, 3D-printed, single-photon head-mounted fluorescent microscopes that simultaneously record audio and video (Fig. 2). These microscopes enabled us to record 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 analogue 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 five electrical wires run out from the camera: one wire each for camera power, ground, audio, NTSC analogue 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 analogue 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 anaesthetized and head fixed. The miniaturized microscope was held using 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.

Histological verification of genetic tool properties

DiI was injected into area X as described above. Three days later, ~800 nl lentivirus was injected into HVC using the DiI demarcation. In finches, this virus infected predominately PNs[46]. In this project we analysed neurons with sparse activity that do not match the tonic activity of interneurons in HVC. The virus was injected into four sites, at least 0.2 mm apart and at two depths (matching the in-vivo imaging experiment's procedure above). About four weeks later, the bird was euthanized (by intracoelomic injection of 0.2 ml 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. 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 using a Nikon Eclipse Ni-E tabletop microscope (Extended Data Fig. 4a).

Expression specificity to excitatory neurons. The fixed tissue was immersed in 20% sucrose solution overnight and then 30% sucrose solution over the following night, frozen and sectioned into 30-μm sagittal slices (Cryostat, Leica CM3050S). Following work in zebra finches[47], the slices were stained using antibodies against the calcium binding interneuron markers calbindin (1:4,000, SWANT), calretinin (1:15,000, SWANT), and parvalbumin (1:1,000, SWANT) by overnight incubation with the primary antibody at 4 °C and with a secondary antibody (coupled to Alexa Fluor 647) for 2 h 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 3-μm-thick sections through the tissue (Extended Data Fig. 4b). The images were inspected for co-stained cells (for example, see Supplementary Videos 1–7). The results ruled out any co-expression of GCaMP and calbindin or calretinin. We found two cells that expressed both parvalbumin and GCaMP (Supplementary Video 5 shows one example; <0.5% of parvalbumin-stained cells, <0.01% of GCaMP-expressing cells), possibly replicating a previous observation of parvalbumin expression in HVC PNs[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 were copious singers (≥50 songs per day) and produced at least 10 different types of syllable.

Video and audio recording. All data used in this manuscript were acquired between late February and early July—a period during which canaries perform their mating season songs. To avoid overexposure of the fluorescent proteins, data collection was done during the morning hours (from sunrise until about 10 am) and the daily accumulated LED-on time rarely exceeded 30 min. 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 the detection of song periods markedly. Data files from the first couple of weeks, a period during which the microscope focusing took place and the birds sang very little, were not used. Additionally, data files from (extremely rare) days on which video files were corrupted because of tethering malfunctions were not used.

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/extract-media) along with the accompanying timestamps. Video frames were interpolated in time and aligned to an average frame rate of 30 Hz. Audio samples were aligned and trimmed in sync with the interpolated frame timestamps. To remove out-of-focus bulk fluorescence from the 3D 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 3D video data, $V(x,y,t)$.

Audio processing. Song syllables were segmented and annotated by a semi-automatic process. First, 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 labelled set was then used to train a deep learning algorithm ('TweetyNet') 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 were fed to TweetyNet in segments of 1 s and the output of TweetyNet was the most likely label for each 2.7-ms time bin in the recording.

Assuring the separation of syllable classes. To make sure that the syllable classes were 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 syllable acoustics. Apart from the clear between-class separation of different syllables for syllables that precede complex transitions, we checked the within-class distinction between contexts that affect the transition. To do that, we used previously published parameters[48] and treated each syllable rendition as a point in an eight-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:

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

Where $\mu_A - \mu_B$ is the L_2 distance between the centres of the distributions and σ_A^2 and σ_B^2 are the within-group distance variances from the centres. Extended Data Figure 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 extracts the probabilistic suffix tree[1] (PST) for each transition (https://github.com/jmarkow/pst). In brief, the tree is a directed graph in which each phrase type is a root node that represents 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 that rarely deviated from simple ellipses. Notably, 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, $f(t) = \sum_{(x,y)\in \mathrm{ROI}} V(x,y,t)$, 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 denoised fluorescence, $(\Delta f/f_0)_{\text{denoised}}$, was estimated from the relative fluorescence change using previously published modelling of the calcium concentration dynamics and the added noise process caused by the fluorescence measurement[49].

Seeking ROIs with sequence correlations. As each ROI was sparsely active in very few phrase types, we first sought ROIs that were active during a phrase type and then tested whether it showed correlations to preceding or following phrase identities. We used the following two-step 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}, was taken as a measure of within-phrase peak values to reduce outliers. Irrespective of the phrase boundaries, periods of time during which an ROI was active were separated from baseline noise fluctuations by fitting the signal within an ROI, $s(t)$, with a two-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, $\mathbf{\Theta}_t = (\mu_t,\sigma_t)$, 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 first-order time-independent Markov transition probabilities, $R = p(\mathbf{\Theta}_{t+1}|\mathbf{\Theta}_t)$, a 2×2 matrix of transition probabilities between two states ('activity' and 'noise'). To estimate the sequence of states (the hidden process $\mathbf{\Theta}$), we maximize the log-likelihood: $L\{\mathbf{s},\mathbf{\Theta},R,\boldsymbol{\mu},\boldsymbol{\sigma}\} = \sum_t \log p(\mathbf{\Theta}_t|\mathbf{\Theta}_{t-1}) + \sum_t \log p(s(t)|\mathbf{\Theta}_t)$. 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: $\Delta ff_{90} > 0.1$ (that is, fluorescence fluctuation is larger than a 10% deviation from baseline); and $\text{HMM}_P > 0.1$ (that is, the phrase type carries neural activity in 10% of occurrences or more). In our data set, this threshold is roughly equivalent to ignoring ROIs that are active only once or twice during a recording day.

Step 2: test sequence correlations. First-order relationships between the signal integral (summed across time bins in the phrase) and the upstream or downstream phrase identities were tested using a one-way ANOVA. The entire set of songs for each bird was used to calculate the first-order phrase transition probabilities, $P_{ab} = P(a \to b)$, for all phrases a and b. Second-order relationships were tested between the signal integral and the identity of the second upstream (downstream) phrase identity for all intermediate phrase types that preceded (followed) the phrase-in-focus in at least 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 that led to the significant ANOVA contained only one song. Data used for ANOVA tests are represented in Extended Data figures by box plots marking the median (centre line); upper and lower quartiles (box limits); extreme values (whiskers), and outliers (+ markers).

The data were not tested for normality before 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[50] 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)." In addition, carrying tests for normality will create a bias in our analyses. Each neuron that is tested for phrase sequence correlation is recorded in a different number of songs. Testing for normality will create a bias towards larger numbers of songs and against high-order correlations.

Nevertheless, we repeated the analyses in this manuscript with non-parametric one-way ANOVA (Kruskal–Wallis). Although ~15% fewer neurons passed the more stringent tests, all the results in this article remained the same. We include a summary of the non-parametric statistics as Supplementary Note 2.

Note that, in this procedure, sparsely active ROIs or ROIs that were active in rare phrase types were not tested for sequence correlation. In the main text we reported that 21.2% of the entire set of ROIs showed sequence correlation. This percentage includes 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 first- and second-order correlations).

Phrase specificity. The fraction of phrase repetitions during which a ROI is 'active', HMM_P, was also used to calculate the phrase specificity of an ROI (Fig. 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. Then, the number of phrase types that accounted for 90% of the ROI's activity was calculated.

Transition-locked activity onsets. The hidden Markov modelling of neural activity was used to identify signal onsets at transition from the 'noise' to the 'active' states (Fig. 2e, Extended Data Fig. 7d). 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 for which the sequence-correlated activity initiated during the phrase transition in the majority of cases were suspected as transition-locked representations. These activity rasters were manually examined and a small number of representations (nine) were excluded from population-level statistics because they appeared reliably and exclusively in specific transitions. Signals that occur exclusively in specific transitions are trivially sequence correlated but simply reflect the ongoing behaviour. This exclusion does not change the results in this paper.

Controlling for phrase durations and time-in-song confounds. In songs that contain a fixed phrase sequence, as in Fig. 2d, we calculated the significance of the relation between $s = \sum_{t\in P}(\Delta f/f_0)_{\text{denoised}}$, an integral of the signal during one phrase in the sequence, the target phrase P, and the identity of an upstream (or downstream) phrase that changes from song to song using a one-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; and 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 a one-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. In Fig. 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 $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ 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 $\mathrm{CI} = \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.

In this comparison there is no bias from the conditions of the statistical test (one-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 fewer) song contexts are not more likely to reach statistical significance by chance.

Contrasting the strength of sequence correlation to past and future events. For one-way ANOVA tests, we estimated the significance of the difference in η^2-statistics (fraction explained variance) calculated in past versus future correlations using the following bootstrapping procedure. First, we pooled all η^2-statistics together. Then we randomly split the pool into 'past' and 'future' groups of the same size as the data in Fig. 2e and calculated the mean value in each group. We repeated this process 1,000,000 times and used this bootstrapped distribution to calculate a P value for the original difference between means. This process was carried out separately for first-order sequence correlations and for second-order or greater sequence correlations (Extended Data Fig. 6i).

Peak location, onset location, and relative duration of sequence correlated activity. The data in Fig. 3a were used to create the following three distributions (Fig. 3d). 1, Relative peak timing: the trial-averaged signals (rows in Fig. 3a differ in ROIs and phrase type) were calculated after time-warping the signals to a fixed phrase duration, $T_{\mathrm{phrase}} = 1$, the onset of which is set to $T_{\mathrm{onset}} = 0$. The timing of the signal peak, t_{peak}, is therefore already normalized because $t_{\mathrm{peak}} = (t_{\mathrm{peak}} - T_{\mathrm{onset}})/T_{\mathrm{phrase}}$.

2, Relative onset timing: the signal in each trial that contributed to Fig. 3a was fitted with a hidden Markov model (as explained in 'Seeking ROIs with sequence correlations'). The onset time point of the signal state, t_{onset}, was normalized with respect to the phrase onset time, T_{onset}, and the phrase duration, T_{phrase}:

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

3, Relative signal duration: a threshold at 0.5 was 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. To simulate the expected calcium indicator signal in response to a spike train, $sp(t)$ (Extended Data Fig. 7a), we used the empirical single-spike response:

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

Corresponding to a rise time constant of 45 ms and a decay time constant of 142 ms (see supplementary table 3 in ref. [25]). The above kernel is a low boundary on the rise time because it assumes 45 ms 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:

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

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

Testing whether sequence-correlated neurons prefer one or more song contexts. For neurons with significant sequence correlations (one-way ANOVA, described above), we used Tukey's post hoc analysis to determine whether this sequence correlation resulted from a significant single preferred context or significant several preferred contexts (Extended Data Fig. 9). 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 $P < 0.001$). As the post hoc test uses a subset of the songs, it is weaker than the one-way ANOVA, and some neurons do not show a clear preference to one context or more but still have sequence correlation (grey in Extended Data Fig. 9f).

Maximum fluorescence images for comparing context-dependent signals. For songs that contain a fixed phrase sequence and a variable context element, such as a preceding phrase identity, maximum projection images were created, as above, but using only video frames from the target phrase (for example, the pink phrase in Fig. 2d). Then, the sets of maximum projection images in each context (for example, identity of upstream phrase) were averaged, assigned orthogonal colour maps (for example, red and cyan in Extended Data Fig. 5) and overlaid. Consequentially, regions of the imaging plane that have no sequence preference would be closer to grey scale, whereas ROIs with sequence preference would be coloured. In Extended Data Figs. 5, 9, we used a sigmoidal transform of the colour saturation to amplify the contrast between colour and grey 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 were subtracted and pseudo-coloured (insets in Extended Data Fig. 5).

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

Label prediction from clustered network states

The signal integral during a target phrase (pink in Extended Data Fig. 10a) was used to create network states—vectors, composed of signals from four jointly recorded ROIs. The averages of the vectors, belonging to the contexts defined by the first upstream (or downstream) phrase label define label-centroids. Then, labels of individual songs were assigned to the nearest neighbouring centroid (Euclidean).

Bootstrapping mutual information in limited song numbers

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

We sub-sampled three out of four ROIs in each permutation and converted their signal to binary values by thresholding the signal integral. Next, we reduced the number of phrase labels by merging. Specifically, in Extended Data Fig. 10, the least common label in downstream states was randomly merged with one of the other labels. In the upstream labels, the least common label was merged after a random division of the other four labels, to form two groups of two.

The mutual information measures were then calculated for each of the 48 possible state spaces and divided by the entropy of the behaviour state, leading to the scatter shown in Extended Data Fig. 10b. The margin of error was estimated from the standard deviation. The 0.95 quantile level of the null hypothesis was created by randomly shuffling each variable to create 1,000 surrogate datasets and repeating the measures. The shuffled set was used to create a sample distribution and to calculate the significance of the differences in Extended Data Fig. 10b using a *z*-test with the sample mean and standard deviation.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data can be found at figshare (https://figshare.com/) with https://doi.org/10.6084/m9.figshare.12006657. Source data are provided with this paper.

Code availability

All custom-made code in this manuscript is publicly available in Github repositories (https://github.com/gardner-lab/FinchScope; https://github.com/gardner-lab/video-capture; https://github.com/gardner-lab/FinchScope/tree/master/Analysis%20Pipeline/extract-media; https://github.com/yardencsGitHub/BirdSongBout/tree/master/helpers/GUI; https://github.com/yardencsGitHub/tweetynet; and https://github.com/jmarkow/pst).

45. Stokes, T. M., Leonard, C. M. & Nottebohm, F. The telencephalon, diencephalon, and mesencephalon of the canary, *Serinus canaria*, in stereotaxic coordinates. *J. Comp. Neurol.* **156**, 337–374 (1974).
46. Liberti, W. A., III et al. Unstable neurons underlie a stable learned behavior. *Nat. Neurosci.* **19**, 1665–1671 (2016).
47. Wild, J. M., Williams, M. N., Howie, G. J. & Mooney, R. Calcium-binding proteins define interneurons in HVC of the zebra finch (*Taeniopygia guttata*). *J. Comp. Neurol.* **483**, 76–90 (2005).
48. Wohlgemuth, M. J., Sober, S. J. & Brainard, M. S. Linked control of syllable sequence and phonology in birdsong. *J. Neurosci.* **30**, 12936–12949 (2010).
49. Zhou, P. et al. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *eLife* **7**, e28728 (2018).
50. Howell, D. C. *Statistical Methods for Psychology* (Cengage Learning, 2009).

Acknowledgements This study was supported by NIH grants R01NS089679, R01NS104925, R24NS098536 (T.J.G.) and R24HL123828, U01TR001810 (D.N.K.) We thank J. Markowitz, I. Davison, and J. Gavornik for discussions and comments on this manuscript, and Nvidia Corporation for a technology grant (Y.C.).

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.N.K. produced lentivirus. Y.C. and J.S. designed surgical procedures. Y.C., J.S., and D.S. performed animal surgeries. Y.C. and D.P.L. built the experimental setup. Y.C. and J.S. gathered the data. Y.C. and D.S. performed histology and immunohistochemistry. Y.C. designed and wrote the machine-learning audio segmentation and annotation algorithm. Y.C. analysed the data. Y.C., W.A.L.III, L.N.P., and T.J.G. wrote the manuscript.

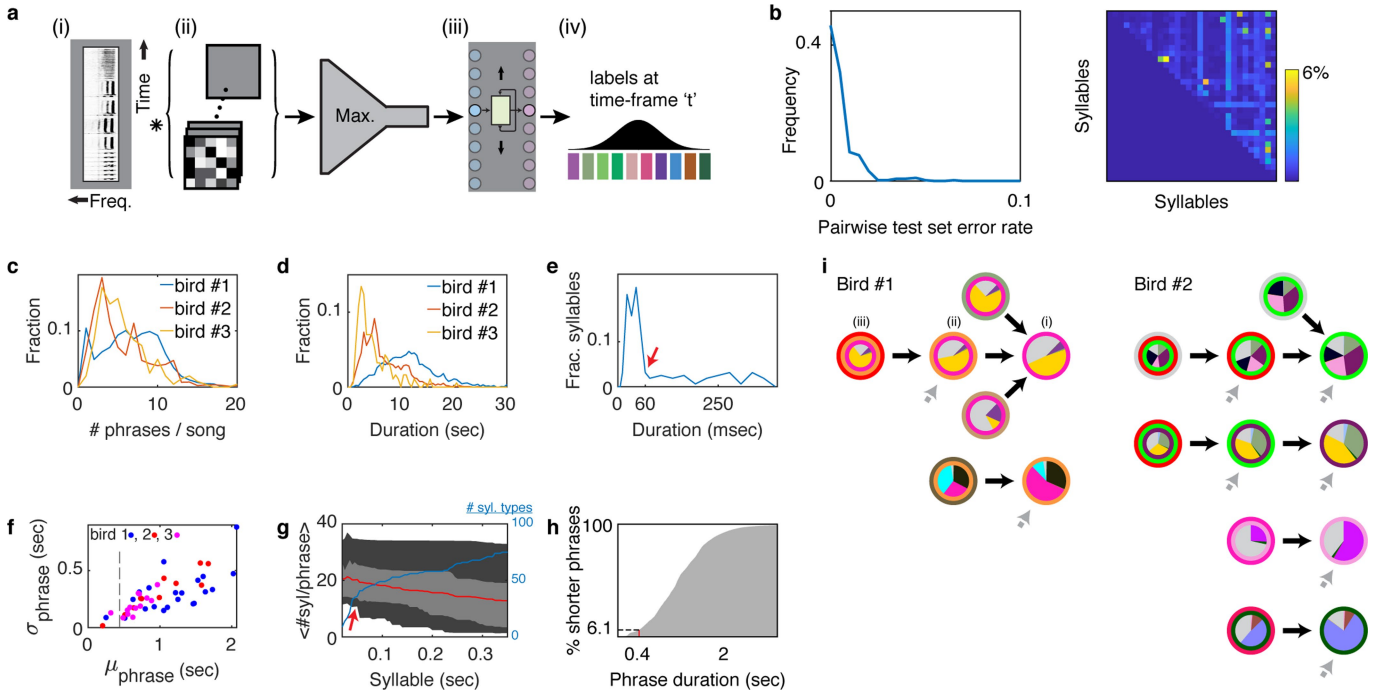
Competing interests The authors declare no competing interests.

Additional information

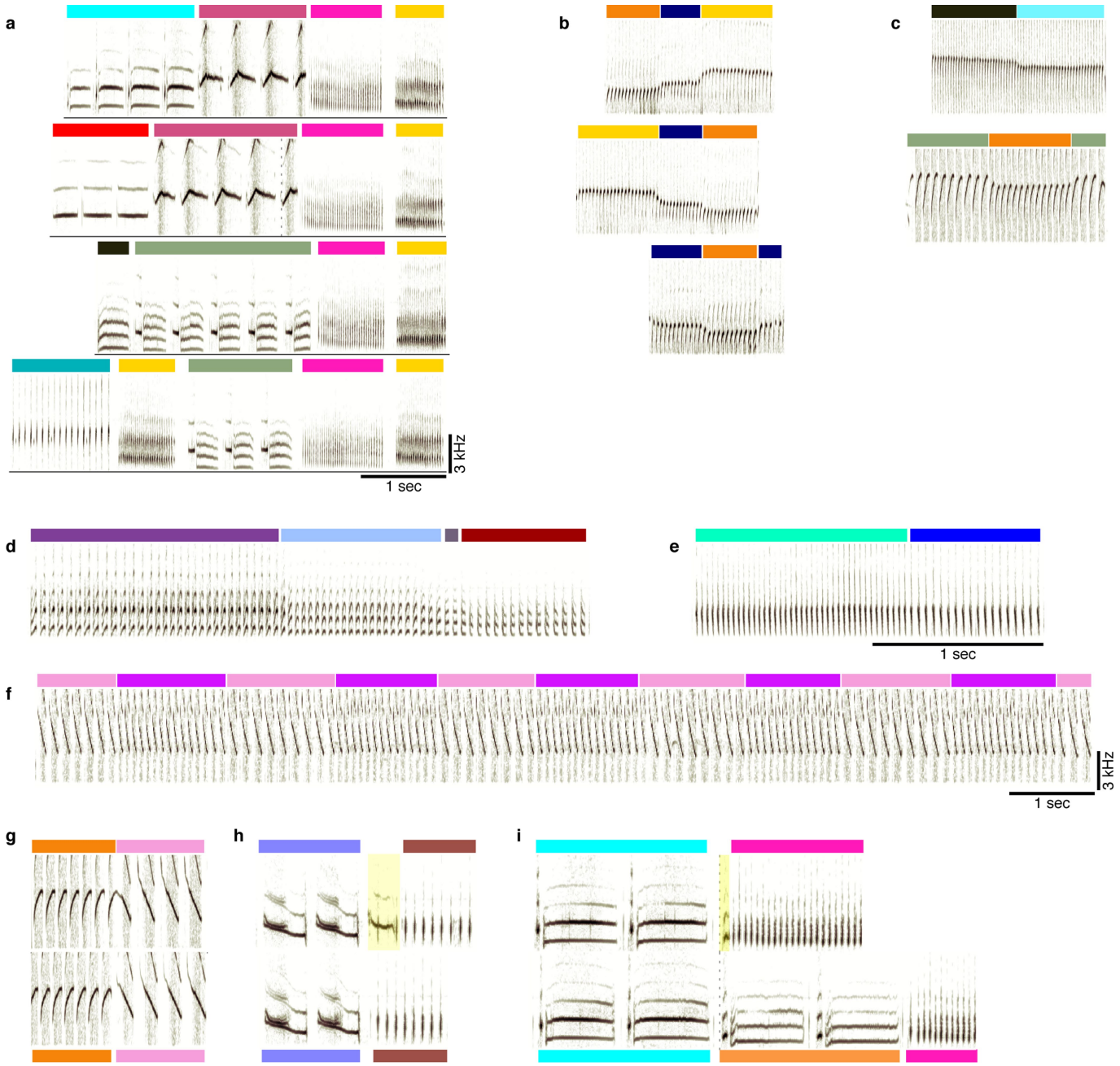
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2397-3.

Correspondence and requests for materials should be addressed to Y.C. or T.J.G.

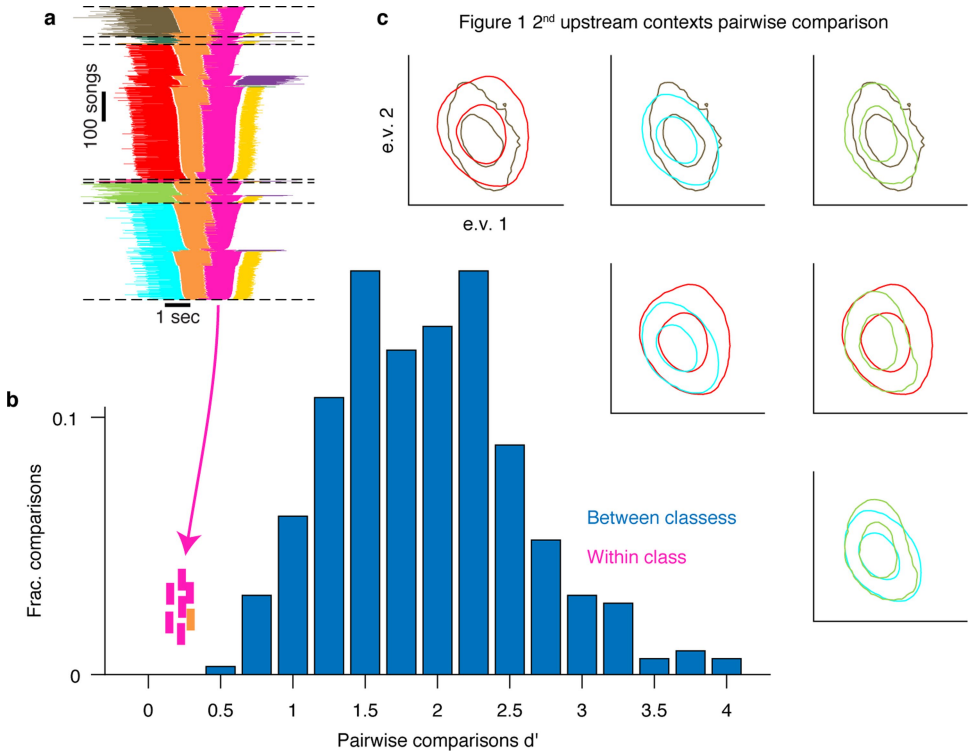
Peer review information *Nature* thanks Jesse Goldberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



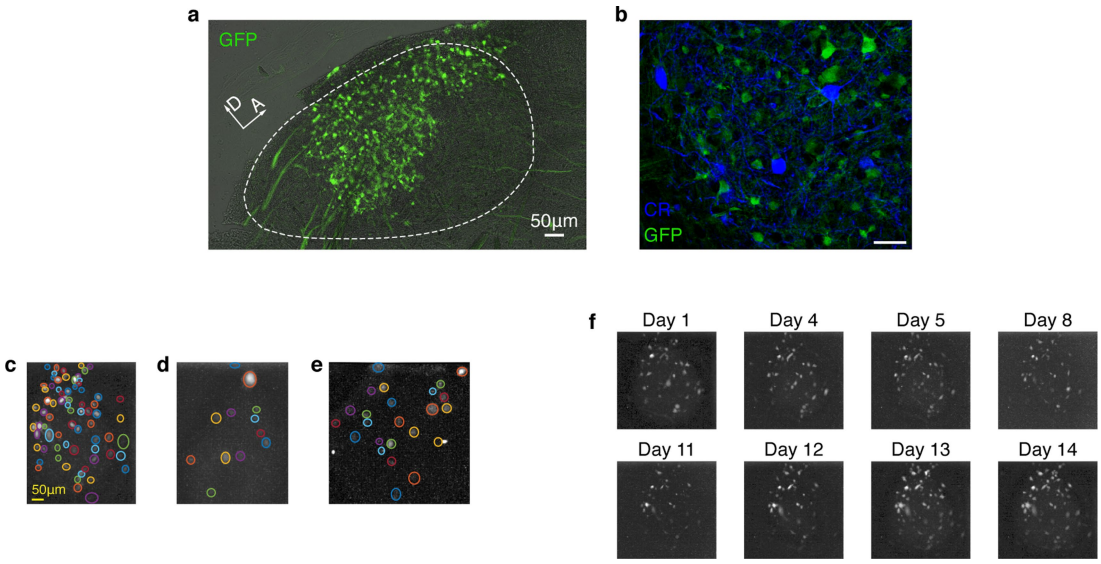
Extended Data Fig. 1 | Canary song annotation and sequence statistics. **a**, Architecture of syllable segmentation and annotation machine learning algorithm. (i) A spectrogram is fed into the algorithm as a 2D matrix in segments of 1 s. (ii) Convolutional and max-pooling layers learn local spectral and temporal filters. (iii) Bidirectional recurrent LSTM layer learns temporal sequencing features. (iv) Projection onto syllable classes assigns a probability for each 2.7-ms time bin and syllable. **b**, After manual proofreading (see Methods), a support vector machine classifier was used to assess the pairwise confusion between all syllable classes of bird 1 (see 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%. As the higher values occurred only in phrases with 10 s 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 of making a mistake in identifying a phrase type is negligible. **c**, Number of phrases per song for the three birds used in this study. **d**, Song durations for the three birds. **e**, Mean syllable durations for 85 syllable classes from three birds. Red arrow marks the duration below which all trill types have more than ten repetitions on average. **f**, Relation between phrase class mean duration (*x* axis) and standard deviation (*y* axis). Syllable classes (dots) of three birds are coloured according to bird number. Dashed line marks 450 ms (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 grey marks the 25% and 75% quantiles and dark grey marks the 5% and 95% quantiles (blue line marks the number of syllable types contributing to these statistics). The red arrow matches the arrow in **e**. **h**, Cumulative histogram of trill phrase durations. **i**, All complex phrase transitions with second-order or higher dependence on song history context (for birds 1 and 2). For each phrase type that precedes a complex transition, the context dependence is visualized by a PST (see Methods). Transition outcome probabilities are marked by pie charts at the centre of each node. The song context (phrase sequence) that leads to the transition is marked by concentric circles, the innermost being the phrase type that preceded 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 (for example, i–iii for first- to third-order Markov chains). Grey arrows indicate additional incoming links that are omitted for simplicity.



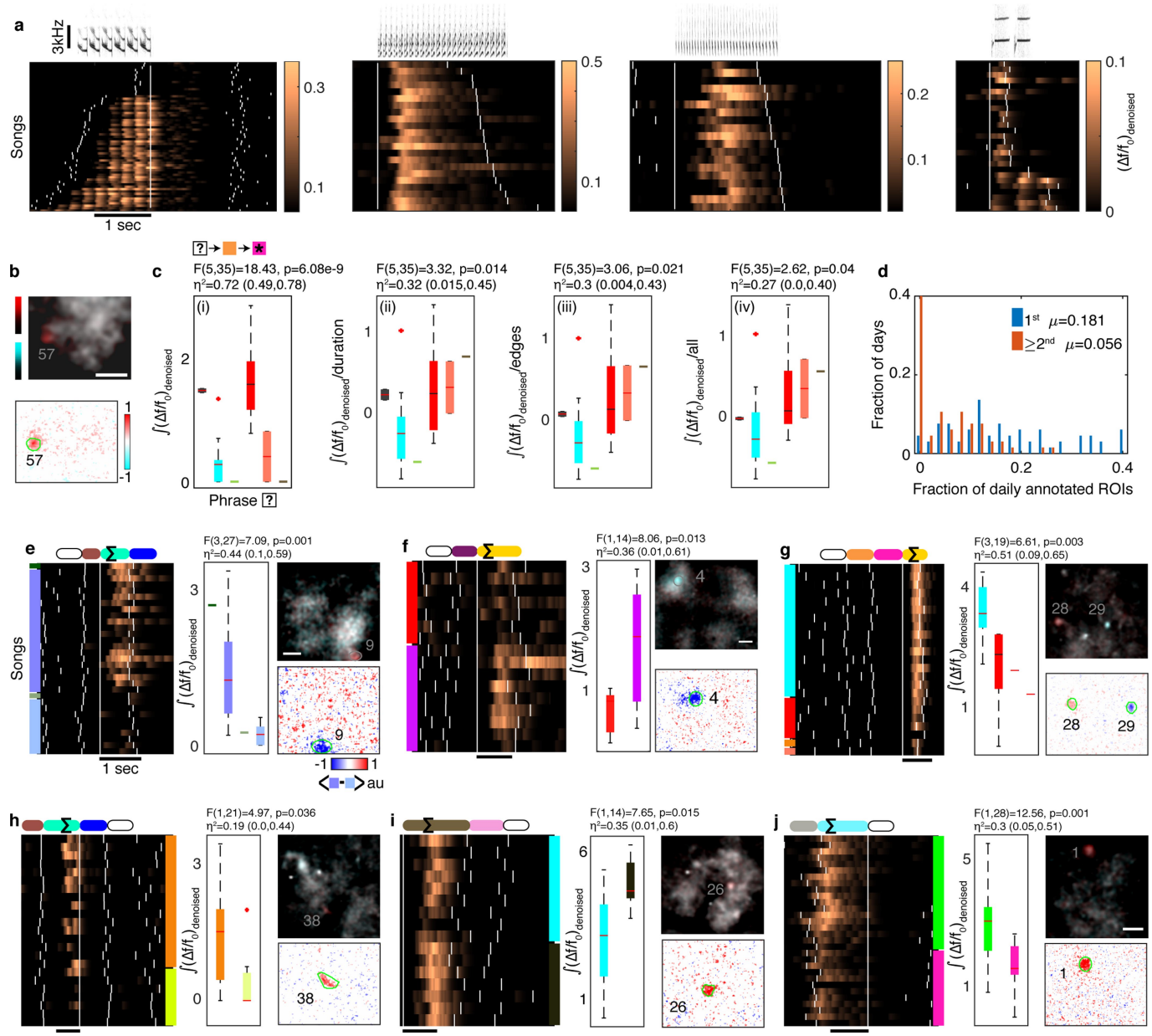
Extended Data Fig. 2 | Examples of canary song phrase sequences, rare inter-phrase gaps, and aberrant syllables. a, Additional spectrograms of phrase sequences (colours above the spectrograms indicate phrase identity) that lead to a repeating pair of phrases (pink and yellow). **b**, Examples of flexible phrase sequencing comprising 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 shows a change in inter-syllable gaps. **f**, Bird 2 shows 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 at the beginning of the second phrase (highlighted) that, in **i**, resembles the onset of an orange phrase type.



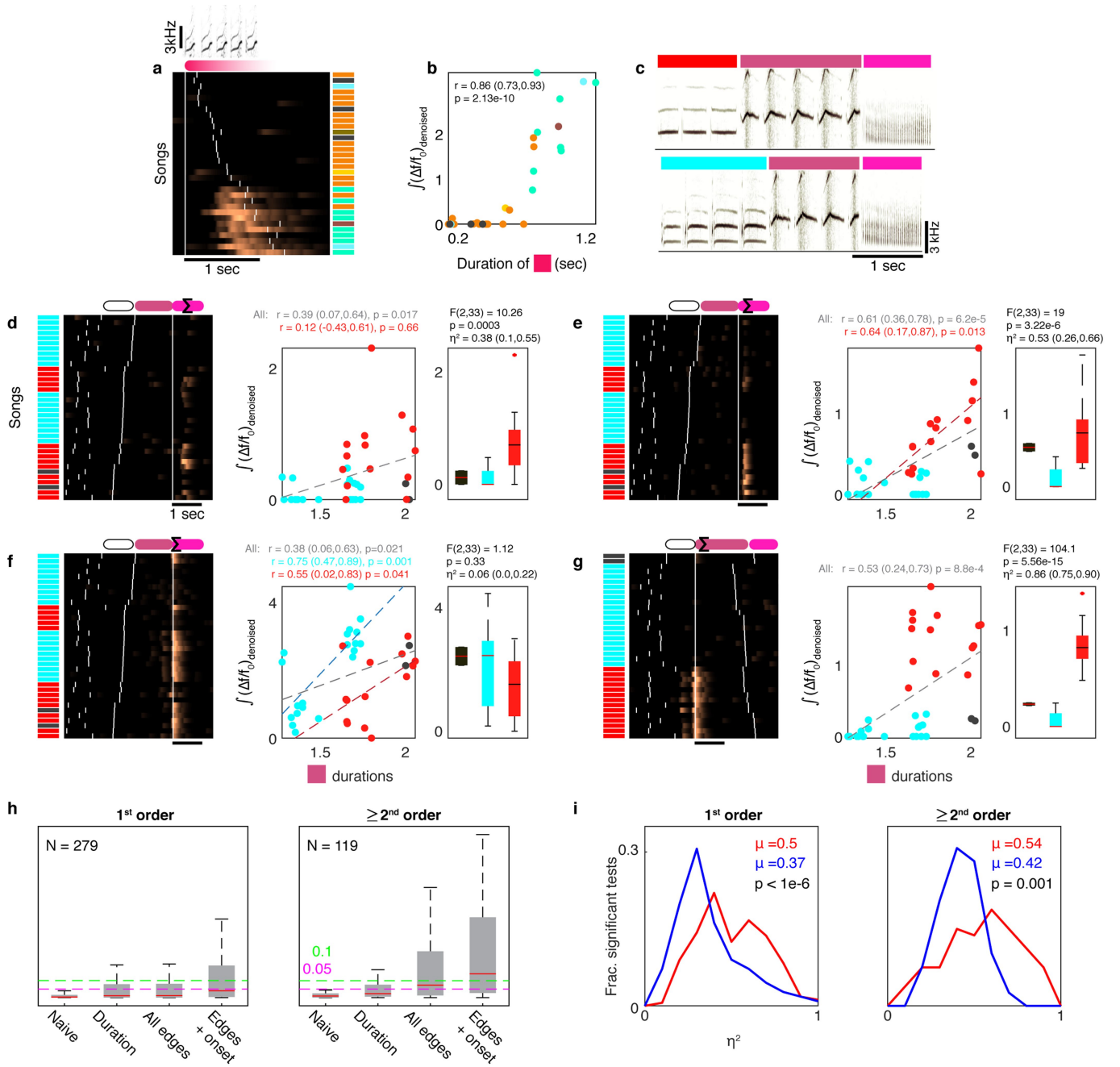
Extended Data Fig. 3 | An example in which the context-dependence of syllable acoustics before complex transitions is too small for clear distinction. a, Same as Fig. 1b. A summary of all phrase sequences that contain a common transition reveals that the choice of what to sing after the pink phrase depends 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 first phrase, the identity of the last phrase, and then the duration of the centre phrases. **b**, The discriminability (d', x axis) measures the acoustic distance between pairs of syllable classes in units of the within-class standard deviation (see Methods). Bars show the histogram across all pairs of syllables identified by human observers (see Methods), corresponding to about 99% or more identification success (Extended Data Fig. 1b). The pink ticks mark the d' values for six within-class comparison of the main four contexts in **a**. The orange tick marks the d' for 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 **b**. Each inset shows overlays of two distributions marked by contours at the 0.1 and 0.5 values of the peak and coloured according to context in **a**. The distributions are projected onto the two leading principle components of the acoustic features (see Methods, in the space defined by eight acoustic features[48]). While some of these distributions are statistically distinct, they allow for only about 70% context identification success in the most distinct case.



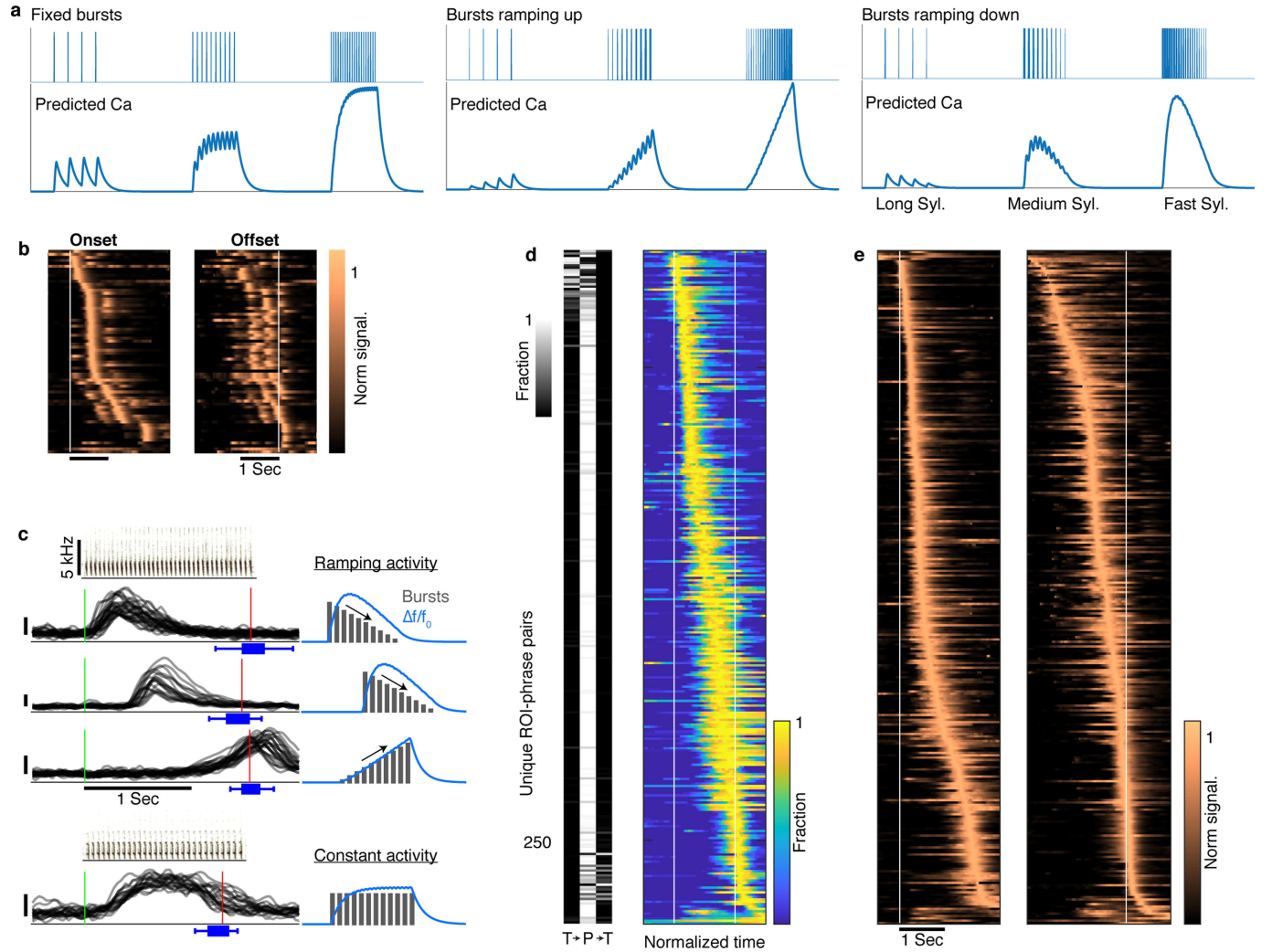
Extended Data Fig. 4 | Calcium indicator is expressed exclusively in HVC excitatory neurons and imaged in annotated ROIs. a, Sagittal slice of HVC showing GCaMP-expressing PNs (experiment repeated in five birds with similar results). **b**, We observed no overlap between transduced GCaMP6f-expressing neurons and neurons stained for the inhibitory neurons markers calretinin (CR), calbindin, and parvalbumin (calretinin stain shown, staining experiment repeated six times for each marker with similar results). **c**–**e**, Examples of daily ROI annotation in three birds (1–3). Coloured circles mark different ROIs, manually annotated on maximum fluorescence projection images on an exemplary day (see Methods). **f**, Maximum fluorescence images (from bird 1; see Methods) revealing fluorescence sources, including sparsely active cells, in the imaging window across multiple days.



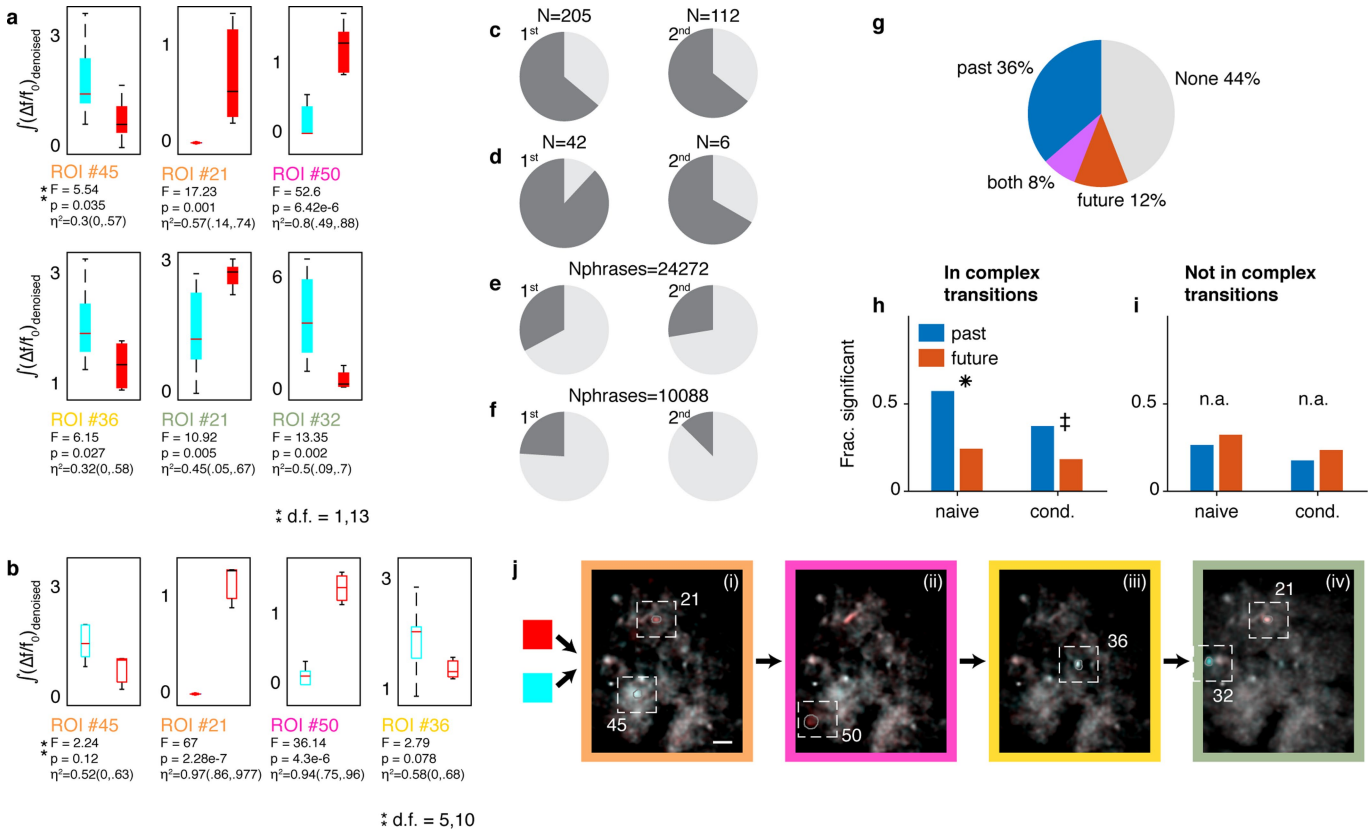
Extended Data Fig. 5 | Syllable and phrase-sequence-correlated ROIs from three birds. a, Sonograms above rasters from four ROIs from three birds. White ticks indicate phrase onsets. The fluorescent calcium indicator is able to resolve individual long syllables. **b**, Top, average maximum fluorescence images during the pink phrase in Fig. 2d (compare the two most common contexts in orthogonal colours (red and cyan)). Scale bar, is 50 μm. Bottom, difference of the overlaid images. ROI outlined in green. **c**, (i) One-way ANOVA (F, P, η^2 and its 95% CI) tests the effect of contexts (x axis, second preceding phrase type in $n = 41$ sequences) on the signal (y axis) during the target phrase (marked by star) in Fig. 2d. Lines, boxes, whiskers, and plus symbols show the median, first and third quartiles, full range, and outliers. (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. Colours correspond to phrases in Fig. 2d. **d**, Fractions of daily annotated ROIs showing sequence correlation in all three 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 Σ) is strongly related to non-adjacent phrase identities (empty lozenges in colour-coded phrase sequence). Songs are arranged by the phrase sequence context (left or right colour patches for past and future phrase types, respectively). White ticks indicate phrase onsets. Box plots and contrast images as defined in **b**, **c**. $n = 31, 16, 23, 23, 16$ and 30 songs contribute to **e**–**j**, respectively. **e**, **f**, Similar to main Fig. 2d, $(\Delta f/f_0)_{denoised}$ from ROIs with second-order upstream sequence (colour coded) from two more birds. **g**, Third-order upstream relation. **h**, **i**, Second-order downstream relations. **j**, First-order downstream relation from another bird.



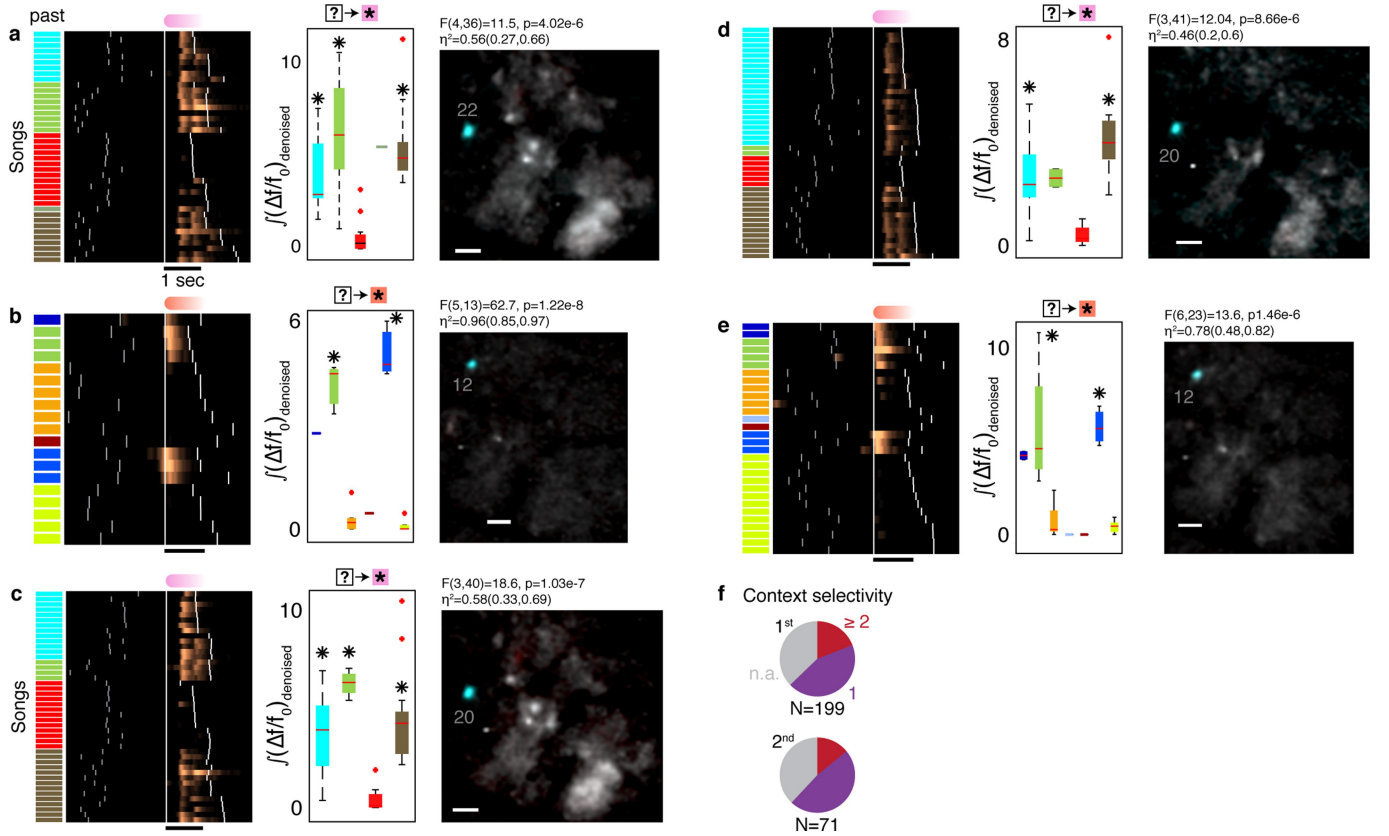
Extended Data Fig. 6 | Durations and onset times of phrases also correlate with their sequence, but cannot fully account for HVC activity. **a**, $(\Delta f/f_0)_{denoised}$ signal traces (ROI 18, bird 3) during one phrase type (red) arranged by phrase duration. Coloured barcode annotates the final phrase in the sequence. **b**, The signal correlates to the red phrase's duration (r (95% CI), P: Two-sided Pearson's test for $n = 32$ songs). Colours match barcode in **a**. **c**, Sonograms of two phrase sequences. **d**–**g**, ROI signals during $n = 36$ sequences containing the last two phrases in **c** have various relations to the duration of the middle (purple) phrase (middle; scatter plots as in **b**, dashed lines indicate significant correlations) and the identity of the first phrase (right; colours, one-way ANOVA (F, P, η^2 (95% CI)) tests the effect on the signal Σ. Whiskers, boxes, and lines show full range, first and third quartiles, and medians, respectively). **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 **a**, but the signal is in the second phrase, not the third. **h**, Distributions of one-way ANOVA P values (y axis; whiskers, boxes, and red lines show full range, first and third quartiles, and medians, respectively) relating phrase identity and signal for adjacent phrases ($n = 279$ independent first-order tests, left) and non-adjacent phrases ($n = 119$ independent second- or higher-order tests, right). Tests were 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). Coloured dashed lines mark $P = 0.05$ and 0.1. **i**, Effect size (η^2 denotes fraction of variance accounted for by the signals' context dependence) of past (red) and future (blue) one-way ANOVA tests for first-order (left, $N = 279$ tests) and second- or higher-order (right, $n = 119$) correlations. The difference in the mean value (μ) is tested using one-sided bootstrap shuffles (P values, see Methods).



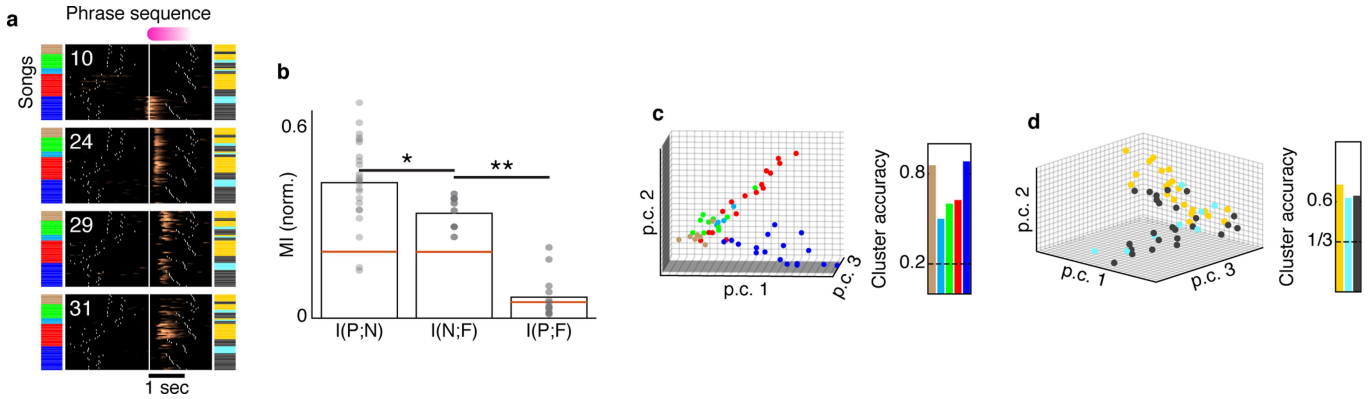
Extended Data Fig. 7 | Signal shape and onset time of sequence-correlated HVC neuron activity reflect within-phrase timing. a, Simulation of calcium indicator (GCaMP6f) fluorescence corresponding to syllable-locked spike bursts in HVC PNs. Syllable-locked spike bursts are convolved with the indicator's kernel (see 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 (400–500ms), medium-length syllables (100 ms) and short syllables (50 ms). **b**, Complementing Fig. 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 have properties of within-phrase timing codes. Example raw $\Delta f/f_0$ signals (y axis, 0.1 marked by vertical bar) of four ROIs aligned to the 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 (bottom). **d**, Left, barcodes show the fraction of signal onsets found in the preceding transition, within the phrase, and in the following transition (T→P→T, see Methods). Rows correspond to the phrases in Fig. 3a. Right, rows show the average signal state occupancy estimated from HMMs fitted to the single-trial data used for Fig. 3a. The resulting traces are time-warped to fixed phrase edges (white lines). **e**, Single-trial data from Fig. 3a 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 first and third quartiles, and lines show medians) for ROIs in Fig. 4a. Text label is colour coded by phrase type in i–iv. F numbers, P values, and η^2 (95% CI) for one-way ANOVA relating history (x axis) and signal (y axis) in $n = 15$ song sequences. **b**, ROIs in **a** retain their song-context bias for songs that terminate at end of the third phrase rather than continuing. Box plots repeat the ANOVA tests in **a** for $n = 16$ songs in which the last phrase is replaced by the end of the song. **c**–**f**, Dark grey slices indicate the fraction of correlations that occur in complex behavioural transitions. **c**, **d**, Data from Fig. 4c separated into the two birds. **e**, **f**, The fraction in **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 effects of the preceding and following phrase types on the signal (see Methods). Pie chart shows the percentage of sequence-correlated ROIs that were 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 with the preceding phrase type (blue) than with the following one (red). This is true in both naive signal values (left, $n = 185$ tests) and after we removed dependencies on phrase durations and time-in-song (right, $n = 185$). One-sided binomial z-test: *proportion difference 0.33 ± 0.09, $Z = 6.45$, $P = 5.5 \times 10^{-11}$; ‡proportion difference 0.19 ± 0.09, $Z = 4.05$, $P = 2 \times 10^{-5}$. **i**, Restricting the analysis to phrase types that are 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**, Fig. 4a showed maximum projection images, calculated with denoised videos (see Methods). The algorithm CNMF-E[49] involves estimating the source ROI shapes, de-convolving spike times and 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 i–iv show the colour bias (cyan or red) that indicates coding of the past phrase with the same colour.



Extended Data Fig. 9 | ROIs that reflect several preceding song contexts. **a**, **b**, ROIs that are active in multiple preceding contexts. $(\Delta f/f_0)_{denoised}$ traces are aligned to a specific phrase onset, arranged by identity of the preceding phrase (colour 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 star) 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 (one-way ANOVA). Asterisks mark contexts that lead to larger mean activity compared to another context (Tukey's multiple comparisons; $n = 41$ songs and $P = 0.01, 7.5 \times 10^{-6}, 5.6 \times 10^{-5}$ in **a**; $n = 19$, $P = 8.8 \times 10^{-7}, 8.15 \times 10^{-8}$ in **b**). Average maximum projection images (see Methods) during the aligned phrase compare the song contexts that lead to significantly higher activity with the other contexts in orthogonal colours (cyan and red for high and low activity, respectively). Scale bar, 50 µm. **c**–**e**, Neurons with similar context preference to those in **a** and **b** on adjacent days. Tukey's multiple comparisons: $n = 44$, $P = 0.001, 4.08 \times 10^{-6}, 1.3 \times 10^{-6}$ in **c**; $n = 45$, $P = 0.0016, 2.85 \times 10^{-6}$ 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 (see Methods). Grey slices (n.a.) mark context-sensitive ROIs for which the post hoc analysis did not isolate a specific context with a larger mean signal. Top (bottom) pie shows selectivity for first (second) preceding phrases.



Extended Data Fig. 10 | HVC neurons can be tuned to complementary preceding contexts. a, Four jointly recorded ROIs exhibit complementary context selectivity. Colour bars indicate phrase identities preceding and following a fixed phrase (pink). For each ROI (rasters), $(\Delta f/f_0)_{\text{denoised}}$ traces are aligned to the onset of the pink phrase (x axis) arranged by the identity of the preceding phrase, by the identity of the following phrase, and finally by the duration of the 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 activity of the four ROIs). Dots, bars, and red lines mark bootstrap assessment shuffles, their means, and the 95% level of the mean in shuffled data (see Methods). *Difference is 0.09 ± 0.03, $Z = 4.3$, $P = 7.3 \times 10^{-6}$; **difference is 0.26 ± 0.02, $Z = 8.9$, $P < 1 \times 10^{-15}$, bootstrapped one-sided z-test. **c**, Signal integrals from the four ROIs in **a** are plotted for each song (dots, $n = 54$ songs) on the three most informative principle components. Dots are coloured 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**, As in **c** but for the first following phrase.

nature research

Corresponding author(s): Yarden Cohen

Last updated by author(s): Mar 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	VOS games' Boom Recorder was used for audio recording from cage microphones. In-house developed software was used for audio and video acquisition in imaging experiments (https://github.com/gardner-lab/video-capture). Histology images were taken with Nikon's 'Elements AR' v4.51.01
Data analysis	Data was analyzed by in-house developed Matlab R2009, R2016b,R2017b (Mathworks) and python (3.6.3) programs. Github links provided in the methods description.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets are available from the corresponding author on request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample sizes are similar to sample sizes used in the field. No statistical methods were used to determine sample size.
Data exclusions	We did not exclude any animal for data analysis. We excluded data from the very rare occasions in which video files were corrupted because of tethering malfunctions.
Replication	We performed recordings from multiple animals to confirm reproducibility. Replications were successful.
Randomization	Our study did not include experimental groups and did not require randomization.
Blinding	Our study did not include experimental groups and did not require blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Immunohistochemistry:
Anti-Calbindin(SWANT CB38, rabbit, lot # 9.03, 1:4000), Anti-Calretinin(SWANT 7697, rabbit, lot # 1893-0114,1:15000), Anti-Parvalbumin (SWANT PV27, rabbit, lot # 2014, 1:1000)

Validation

All antibodies are used in a large number of publications and do not require additional validation.

References in manufacturer's website:
Anti-Calbindin:
Airaksinen M.S., et al, (1997), PNAS 94(4) : 1488-1493
Anti-Calretinin:
1. Schwaller B., Buchwald P., Blümcke I., Celio M.R. and Hunziker W. (1994) Characterization of a polyclonal antiserum against the purified human recombinant calcium-binding protein calretinin. Cell Calcium 14: 639-648.
2. Schiffmann S.N. et al (1999) Impaired motor coordination and Purkinje cell excitability in mice lacking calretinin. PNAS, 96: 5257-5262.
3. Gotzos V., Vogt P. and M.R. Celio (1995) Calretinin is a selective marker for malignant pleural mesotheliomas of the epithelial type. Pathol. Res. Pract. 192:137-147.
4. Doglioni, C. et al. (1996) Calretinin: a novel immunocytochemical marker for mesothelioma. Am. J. Surg. Pathol. 20:1037-1046.
Anti-Parvalbumin:
1. Kretsinger R.H. (1981) Neurosci. Res. Progr. Bull. 19/8, MIT-Press
2. Celio M.R., Heizmann C.W. (1981) Nature 293: 300-302
3. Celio M.R., Heizmann C.W. (1982) Nature 297:504-506
4. Schwaller B., et al. (1999) Am. J. Physiol. 276. C395-403

5. Filice F, Celio M.R., Szabolcsi V. (2017) JCN

References in songbird literature:
1. Wild, J. M., Williams, M. N., Howie, G. J. & Mooney, R. Calcium-binding proteins define interneurons in HVC of the zebra finch (Taeniopygia guttata). Journal of Comparative Neurology 483, 76–90 (2005).
2. Scotto-Lomassese, S., Rochefort, C., Nshdejan, A. & Scharff, C. HVC interneurons are not renewed in adult male zebra finches. European Journal of Neuroscience 25, 1663–1668.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Domestic canaries (Serinus Canaria), American Singer strain, males, older than 1 year.
Wild animals	The study did not involve wild animals
Field-collected samples	The study did not involved samples collected from the field
Ethics oversight	All procedures were approved by the Institutional Animal Care and Use Committee of Boston University (protocol numbers 14-028 and 14-029) with accreditation from the Association for Assessment and Accreditation of Laboratory Animal Care International.

Note that full information on the approval of the study protocol must also be provided in the manuscript.