Supplementary Information

Subjects. Birds were caught on farms in northern Indiana, transported in cages by car to the University of Chicago, and housed in large flight aviaries with ~20 mixed-sex conspecifics until the start of behavioural training. All animals had access to food and water ad libitum while in the flight aviary. Subjects were naive to all the training/testing stimuli at the start of behavioural training.

Stimuli. The organization of starling song can be described hierarchically[1]. Song is typically produced in episodes of continuous singing referred to as 'bouts', with each bout composed of repeated sequences of smaller units called 'motifs'. Each motif, in turn, comprises a stereotyped pattern of several different notes. There are species typical characteristics in the general song structure, and most motifs can be classified into one of four broad species typical categories. However, the structure of different notes in motifs within each category varies greatly and is largely unique to each individual. That is, motif sharing between individuals is relatively rare (~ 1% in birds from the same site[2]). A given adult male may have a repertoire of 50–70 motifs (or more), which he presents in bouts of song ranging from a few seconds to over a minute in length. Depending on its length, a single bout can contain upwards of 20–30 unique motifs, and separate bouts often contain different subsets of the motif types in a given male's repertoire.

We constructed three sets of starling song stimuli for use in this study. Multiple songs were recorded from three adult male starlings (wild caught in Baltimore, MD) implanted with a subcutaneous testosterone Silastic (Dow Corning) pellet. All the vocalizations that each bird produced over a period of three to ten days were digitally recorded (20 kHz sample rate, 16 bit resolution). Given the enormous variability in starling songs, and limitations of stimulus repertoire sizes in electrophysiological experiments, we made no attempt to exhaustively represent the song material of each

bird, and instead relied on a sampling strategy. Five samples of continuous singing from the songs of each bird (15 samples total) were selected, resulting in three sets of songs each containing five samples of a given male's song (Fig. S1). Each song sample was approximately 10 s of continuous singing taken from a single song bout, typically a stretch that avoided the very beginnings and end of a bout. Samples were constrained to start at motif beginnings and end at motif ends, resulting in some variation in sample length. The mean song sample length was 9.72 ± 0.18 s.

We tried to choose song material that was biased toward high within–set motif variety. Typically each song sample contained 4 to 7 different motifs (mean = 5.6), with each motif repeated 2 to 4 times. Motif similarity within the 10-s song samples in a given set was low. The first set of five song stimuli (Fig. S1a) comprised 19 unique motifs (58 motifs total), five of which appeared in exactly two song samples. The second set of five song stimuli (Fig. S1b) comprised 24 unique motifs (56 motifs total), with two motifs appearing in two song samples. The third set (Fig. S1c) comprised 34 unique motifs (65 motifs total), none of which appeared in more than one song sample. There were no common motifs between any of the sets. Because there is no unambiguous definition of a motif, counts may vary slightly depending on the criteria employed. Slight variation in motif counts would affect slightly the absolute number of motifs associated with significant neuronal responses.

We also used two synthetic stimuli: a five-pulse train of a frozen broadband noise (1 s on, 1 s off, 10 s total duration), and a train of three FM cosine waves that each swept from 10 kHz to 10 Hz to 10 kHz over 2 s (2 s ISI, 10 s total duration). The peak power of all stimuli (song samples and artificial stimuli) was normalized.

Behavioural training.

Training stimuli. For each subject, we used song samples from two of the three sets for operant training. Songs from the third set were used as novel stimuli during subsequent electrophysiological testing. The exact pair of stimulus sets used for training was varied across subjects, as was the number of stimuli from each set that the animal was trained to recognize (Table S1). That is, songs that were novel for some subjects, served as half the training stimuli for other subjects, and vice versa. The mean stimulus intensity at a position approximating that of the bird's head during training was ~72 dB SPL

Apparatus. Subjects learned to recognize the training songs using an operant apparatus (Fig.1b), mounted inside a 61 x 96 x 53 cm ID sound attenuation chamber (AC-3, IAC, Bronx, NY). A cage mounted inside the chamber held the subject, while providing access to a 30 x 30 cm operant panel mounted on one side. The panel contained three circular response 'buttons' spaced 6 cm center-to-center, aligned in a row with the center of each button ~14 cm off the floor of the cage, with the entire row centered on the width of the panel. Each response 'button' was a PVC housed opening in the panel fitted with an IR receiver and transmitter that detected when the bird broke the plane of the opening with its beak. This 'poke-hole' allowed starlings to probe with their beaks, a naturally occurring behaviour. Each response opening was illuminated from the rear with an independently controlled LED. Directly below the center button, in the section of cage floor immediately adjacent to the panel, a fourth PVC lined opening provided access to food. A remotely controlled hopper, positioned behind the panel, moved the food into and out of the subject's reach beneath the opening. Acoustic stimuli were transmitted via USB to one of 12 operant stations (one station per animal being trained), converted to analogue via a USB DAC, amplified with an audio amplifier, and then presented to the subject through a small speaker (LCS-1040, Labtec) mounted ~30 cm behind the panel, out of the subject's view. We used custom software

to monitor the subject's responses, and to control the LEDs, food hoppers, chamber light and stimulus presentation according to procedural contingencies.

Operant Procedures. Subjects learned to discriminate and classify the training songs using either a go-nogo or two-alternative choice procedure. In either case, naive subjects were trained initially to work the apparatus thorough a series of shaping routines using food rewards for pecks to lighted response buttons. In all cases, initial shaping occurred within one to two days, and was followed immediately by the start of song recognition training.

For the two-alternative choice (2AC) procedure, subjects initiated a trial by pecking the center response button to trigger immediately the presentation of a single training song stimulus. Following completion of the stimulus, the subject was required to peck either the left or the right response button within 2 s. Half of the stimuli were associated with the left response button and the other half with the right button (Table S1). Correct responses (responses to the correct key, left or right) were rewarded with access to the food hopper for 3 s. Incorrect responses were punished by extinguishing the house light for 2 – 10 s and withholding food. Responses prior to completion of the stimulus were ignored. The trial ended when either the food hopper retracted following a correct response, or the house light re-illuminated following an incorrect response. The inter-trial interval was 2 s. Incorrect responses, and any trial in which the subject failed to peck either the left or the right key within 2 s of stimulus completion initiated a correction-trial sequence, during which the initiating stimulus was repeated on all trials until the animal responded correctly. Correction trial data are not included in the analysis.

For the go-nogo procedure, subjects initiated a trial by pecking the center response button to trigger the immediate presentation of a training song. Following stimulus presentation the animal was required to either peck the center response button again, or

to withhold responses altogether. Responses to half of the stimuli (S+) were reinforced positively with 2-s access to the food hopper. Responses to the other half of the stimuli (S-) were punished by extinguishing the house light for 2–10 s and denying food access. Failure to respond to either S+ or S- stimuli had no operant consequence. Responses prior to completion of the stimulus were ignored. Subjects were initially trained to respond to all stimuli, and only after response rates to all stimuli were high was the differential reinforcement regime introduced. There was no correction trial sequence for the go-nogo procedure.

For both training procedures, the stimulus exemplar presented on any given trial was selected randomly (with replacement) from the pool of all stimuli the animal was learning to recognize (Table S1). The food in the hopper was the same as that provided ad libitum in the aviary. Water was always available. Subjects were on a closed economy during training, with daily sessions lasting from 8 AM until 7 PM. Food intake was monitored daily to insure well-being.

Table S1

	Search/Testing Stimuli			
	Training Stimuli			
Subject	**Class 1**	**Class 2**	**Novel**	**Training regime**
st120	A1-A2-A3-A4-A5	B1-B2-B3-B4-B5	C1-C2-C3-C4-C5	2choice
st125	A1-A2-A3-A4-A5	B1-B2-B3-B4-B5	C1-C2-C3-C4-C5	2choice
st128	A1-A2-A3-A4-A5	B1-B2-B3-B4-B5	C1-C2-C3-C4-C5	2choice
st138	A1-A3-A5	B1-B3-B5	C1-C3-C5	2choice
st150	B2-B4	C1-C2	A1-A2	Go/nogo
st155	B2-B4	C1-C2	A1-A2	Go/nogo
st159	B2-B4	C1-C2	A1-A2	Go/nogo
st163	A1-A2	C1-C2	B2-B4	Go/nogo

Table S1. Stimuli used for operant training and physiological testing of each subject. 'Class 1' and 'Class 2' refer to the two stimulus classes learned during operant training. Class 1 stimuli were associated with the left key in the two choice training and S+ in the go/nogo training. Letters denote the stimulus set from which a given sample was drawn, and numbers denote the specific song sample from that set (Fig. S1).

Electrophysiological testing

Procedure. Once a subject learned to recognize the training songs, on days prior to electrophysiological recordings, we implanted a small pin on the skull under Equithesin anaesthesia (4.5 ml/kg, I.M.). All physiological recordings were conducted with the subject anaesthetized with urethane (20% by volume, 7 ml/kg, I.M.). Subjects were placed in a cloth jacket, comfortably supported, and the head was secured via the pin to a stereotaxic apparatus mounted inside a double-walled sound isolation chamber. Recordings with solder glass-coated Pt/Ir electrodes into either the right or left hemisphere cmHV were made with a dorsal approach through a small (~1 x 1 mm) craniotomy. Extra-cellular waveforms were digitized for offline spike analysis. Most (≈70%) recordings had high S/N and spikes were reliably identified by amplitude (e.g. Fig. 3b of main paper); for the remaining we used more sophisticated spike sorting software[3].

Test Stimuli. Test-stimuli were presented free-field. For all the cells from a given animal, we used the same stimulus ensemble to search and then test each cell's responsiveness. The stimulus composition of the search/test ensemble was dictated by each animal's behavioral training (see Table S1). For example, if the bird was trained to recognize three songs from set 1 and three songs from set 2, then the search/testing ensemble would contain those 6 training songs, three novel songs (from set 3). In addition, the same two synthetic stimuli (10 s duration each; see Text) were presented during searching and testing to each subject. Thus, for this animal there would be a total of 11 stimuli in the search/testing ensemble, 6 of which were familiar to the subject. The familiar songs presented to the subject during testing were the exact same stimuli used for the operant training of that animal (see Table S1).

The stimulus presentation protocols for searching and testing phases were identical. All stimuli in the ensemble were presented randomly without replacement, with an inter-stimulus interval (offset to onset) of 15 to 20 s. During testing, the stimulus presentation was organized into large blocks, with each block comprising 5 or 10 repetitions of each stimulus. The first block of stimuli during the test comprised 5 reps, and the data from the cell were not analyzed if the first block was incomplete (i.e. if unit isolation was lost before all the stimuli in the first block were presented).

Histology. Following completion of the experiment, the subject was deeply anesthetized with an overdose of Nembutal (250 mg /kg), exsanguinated and fixed. The brain was extracted and post-fixed in 10% formalin for several days, then cryoproteced in 30% sucrose prior to coronal sectioning on a freezing microtome. Fifty-micron sections containing cmHV were stained for Nissl using standard histological procedures. We confirmed the position within cmHV of each recording site reported here, by referencing small electrolytic lesions (5-10 μA for 5 s) that marked fiduciary points at or near recording sites. All sites were within 0.35 to 0.75 mm of the midline, 2.3 to 2.7 mm rostral to the caudal bifurcation of the sagittal sinus, between a depth of 1.3 to 2.5 mm.

Data Analysis.

Response strength. We obtained the within-stimulus variance in the spike-rate from the distribution of inverse inter-spike intervals (ISI) occurring during all repetitions of a given stimulus. For spontaneous activity, we used the spikes occurring during a 4-s silent interval beginning 2s after the stimulus ended, then averaged across all stimulus repetitions to yield one value of spontaneous rate variance per cell. The ratio of these two variances was defined as the response strength (RS). RS was typically

calculated for entire 10-s song stimuli, but for one analysis was calculated on a motif-by-motif basis.

Auditory response significance. Auditory responsiveness was assessed quantitatively using the RS values associated with each of the stimuli presented to a given cell. For the 45 cells reported here, the maximum RS value associated with any one stimulus was greater than (1.96 * 1 s.e.) +1, where s.e. is the standard error of the mean RS for all other stimuli presented to that cell, and 1 is the value of RS expected for a non-auditory cell. In addition to the 45 neurons that met this criterion, eight other cells appeared to be auditory but isolation was lost before the criterion for a sufficient data set (5 repetitions per stimulus, see text) was achieved. For 43 other cells, there was either no response to any sound or post hoc analysis revealed that the putative response did not meet statistical significance for any stimulus. Such cells tended to have high S/N and low spontaneous rates. There were many additional apparently non-responsive neurons that we did not catalogue. Thus, the estimate that 50% of cmHV cells were responsive (see text) is an upper bound.

Response strength normalization. When comparing data across cells (e.g. Fig. 2), we standardized the RS values by converting them to z-scores, $Z_i = \left(RS_i - \overline{RS}\right)\big/\sigma$, where RS_i is the response strength of the ith stimulus for a given cell, $\overline{RS}$ is the mean response strength for all stimuli presented to that cell, and σ is the standard deviation of the response strength for all stimuli presented to that cell. Deviations from normality in the underlying distributions were corrected using a square root transformation prior to computing the ANOVA. In all cases, the significance of the results obtained using both the square root transformed and non-transformed data were identical.

SI significance. We assessed the significance of the selectivity index (SI) for each cell by simulating the 'random' response of the cell to all the stimuli actually presented to that cell. For example a given cell might have been tested with six training stimuli

(three from set 1, three from set 2), three novel stimuli (from set 3) and two synthetic stimuli, all of which were repeated ten times. In this case, we would use a Gaussian distributed random variable (mean = 0, variance =1) to generate a 10 x 11 matrix of fictive RS values in which each entry is the 'simulated' RS associated with one repetition of one of the 11 test stimuli, and then find the mean RS across repetitions. This would result in a randomly generated mean RS value for each of the 11 stimuli. These mean RS values were then used to calculate 11 SI scores, one corresponding to each stimulus. This was done 1000 times, saving the 11 SI scores generated on each run, resulting in a random distribution of 11,000 SI scores that matched the experimental conditions for the example cell. The real SI score for this particular cell (calculated from the real RS values associated with each stimulus) was then compared to the value at the 95^{th} percentile of the simulated SI score distribution to assess if the real SI score was likely to have occurred by chance ($p < 0.05$). If the real SI score was greater than the 95^{th} percentile of the simulated SI score distribution, then that cell was termed 'selective' for the stimulus that elicited the maximum RS. Otherwise, the cell was termed 'non-selective' for that stimulus. SI thresholds (95^{th} percentiles) ranged from 0.25 to 0.44.

The mean and variance of the RS model distribution does not effect the resulting SI score distribution. This independence stems from the fact that the SI score is a ratio of RS variances, in which the deviation (away from the mean) of the RS for any given stimulus is expressed as a proportion of the total RS variation across all stimuli. The mean RS is effectively scaled to 0, and the variance to 1, in the equation calculating SI scores.

We also ran the simulations using the empirical distribution of RS scores as a model for RS in the SI simulations, rather than the Gaussian. The results produced with these two models for RS were virtually identical, in both the p-values associated with

the real SI scores (i.e. the location of the actual SI in the simulated distributions; mean p-value with Gaussian model for RS: 0.059±0.010, mean p-value with empirical model for RS: 0.059±0.010; $p = 0.97$, paired t-test), and the value at the 95th percentile of the simulated SI distribution for each cell (i.e. the selectivity threshold; mean for Gaussian model RS: 0.370± 0.011, mean for empirical model RS: 0.371±0.012; $p = 0.72$, paired t-test).

Additional song-selectivity analyses. To explore the pattern of song-selectivity further, we compared the RS derived SI scores to higher and lower selectivity thresholds, and computed SI scores using different dependent measure of response. We derived a higher selectivity threshold by taking only the maximum SI score on each of the 1000 simulation runs, and finding the value at the 95th percentile in the resulting distribution of simulated maximum SI scores. According to this criterion 22/45 cells were selective for one of the test stimuli, and of those cells, 21/22 preferred one of the familiar songs. The proportion of selective cells using this stricter definition was significantly different from chance ($p < 0.005$, chi-square). Even with no threshold for selectivity, that is when one simply takes the stimulus eliciting the maximum RS as being the preferred stimulus, the number of cells that 'prefer' a familiar song (41/45) was significantly different from chance ($X^2 = 16.41$, $p < 0.0005$). We also computed SI scores for each cell using the stimulus-driven spike rate and the raw within-stimulus variance instead of the RS. Based on both these measures, the proportion of selective cells preferring familiar songs (21/23 using spike-rate mean; 24/26 using spike-rate variance) was also significantly different from that expected by chance ($p < 0.01$, both cases, Chi-square). Thus, a broad range of definitions of selectivity yielded results similar to those reported in the text, emphasizing the security of the reported results.

Significance of selective cell proportions. The proportions of cells selective for familiar (training) songs and those selective for unfamiliar songs and artificial stimuli

were compared using the Chi-square statistic. The proportions of familiar and unfamiliar stimuli in the search/test ensemble bias the proportions of cells selective for familiar and unfamiliar stimuli that one expects to observe. We accounted for these biases using the chi-square statistic, which is specifically designed to address the question of whether or not an observed proportion is significantly different than an expected proportion. The expected proportions were computed as follows: for each cell, we took the number of familiar, novel, and synthetic stimuli in that cell's search/test ensemble, and divided each of those three numbers by the total number of stimuli in that ensemble. For example, if six out of 11 stimuli in the ensemble presented to a given cell were familiar, then the expected proportion (i.e. the probability that the cell was selective for a familiar stimulus) was 0.5455. We found the expected proportions for every cell, and then averaged the proportions for each class (familiar, novel, synthetic) across all cells. The averages are the expected proportions of cells selective for familiar, novel, or synthetic stimuli in the population, and are as follows: p(fam) = 0.618, p(nov) = 0.232, p(synth) = 0.15. As an alternative, one might compute the expected proportions of cells selective for familiar, novel, and synthetic stimuli as above, but using the test ensemble distributions for only the selective cells. Based on only the selective cells, the expected proportions are as follows: p(fam) = 0.599, p(nov) = 0.238, p(synth) = 0.162. Comparing these to the observed proportions among the selective cells yields a significant result (X^2= 13.27, $p < 0.005$) similar to that reported in the text.

Spectro-temporal receptive fields. We generated models of the spectro-temporal receptive fields (STRF) for most of the cells in our sample (Fig. S2, a and b) using the algorithm (and software implementation) described by Theunissen et al.[4]. Briefly, we (1) calculated the stimulus auto-correlation and stimulus-response cross correlations, (2) took the Fourier transform along the temporal dimension of both the auto-correlation and cross-correlation, (3) for each temporal frequency, found the eigenvectors and eigenvalues of this spatial auto-correlation matrix of the stimulus, (4)

used the eigenvalues of the spatial auto-correlation function to normalize the cross-correlation in the basis set defined by the eigenvectors, (5) returned to the original spatial basis set, and (6) took the inverse temporal Fourier transform. Because this method is sensitive to the amount of data used to contract the STRF, we report here on data for those cells (N=43) that showed at least one peak in either the frequency or temporal dimension where the gain of the filter (the STRF) exceeded the upper or lower boundary of the 95% confidence interval set by the background noise.

Each STRF provides an estimate of the optimal linear combination of spectro-temporal features that drive a given cell maximally. If a cell's response is linear, then the STRF yields a spectro-temporal representation of the preferred stimulus, and can predict the response to any arbitrary stimulus. To obtain this prediction, we convolved the significant portion of each STRF with the stimulus using a jack-knife routine to obtain predicted firing rates to each test stimulus[4]. The quality of the prediction was assessed by the cross-correlation coefficient (CC) between the PSTH and the predicted spike rate, adjusted for the negative bias from noise. (Fig. S2, c and d). Overall, the STRFs provided relatively poor predictions of a cell's response to a given stimulus (mean CC = 0.285 ± 0.02, range = -0.01 to 0.58). The poor STRF predictions are consistent with non-linear response profiles, and with those observed in the cmHV of other songbird species[4, 5]. Although this analysis provides little insight into the stimulus features that are driving each cell's response, it does render unlikely the possibility that tonotopic or other simple representational topographies can explain the observed responses to conspecific songs in cmHV.

Phasic Response. A neuron's phasic response (*PR*) to each stimulus was quantified as $PR = (D - \sum_{i=1}^{n} \min(x_i, x_{i+1})) \Big/ D$, where D is the duration of the stimulus times the number of stimulus repetitions, x_i is the duration of the i^{th} inter-spike interval (ISI), $\min(x_i, x_{i+1})$ is the minimum interval in each temporally consecutive pair of ISIs, and n

is the last ISI that occurred during all repetitions of the stimulus. When a cell fires tonically and all of the spikes are spaced evenly, the sum of the minimum intervals will be close to *D*, and so PR will be close to zero. When a cell fires more phasically resulting in many small ISIs, the sum will be minimized and PR will approach one. The mean PR for each cell (see Fig. 4) was obtained by averaging PRs to all the stimuli presented to that cell. In theory, this measure of a cell's phasic response may be somewhat corrupted by situations in which a cell has a high stimulus-driven firing rate, and frequent bursts. In practice, this is not likely to significantly affect our data. Cells that have maintained firing rates and frequent bursts would have high PR scores and high response rates. Therefore, a large number of such cells would produce a positive correlation between PR and mean (stimulus-driven) spike-rate. The correlation between PR and spike-rate is significant, but in the opposite direction ($r = -0.685$, $p < 0.0001$, Fischer's r to z transform, see Fig. S3), indicating that the more phasic cells responded with fewer total spikes.

Motif-Feature x Response-Strength Correlations. We used a multiple linear regression to examine potential correlations between acoustic features of the stimulus set, and each neuron's response. We characterized each motif as a set of spectro-temporal features by decomposing the sonogram of every motif using a two-dimensional Haar wavelet transform. To increase the chances that similar features weighted heavily on similar wavelet coefficients, each motif was temporally aligned with other similar motifs (see Figure S1 a–c) and zero-padded to a fixed length prior to decomposition. The transformation of each motif produced a 64 X 4096 matrix of wavelet coefficients, most of which were close to or equal to zero. For a given cell, we averaged the matrices from all the motifs presented to that cell, found the indices for the 50 coefficients with the highest mean, and then extracted those 50 coefficients from the original 64 X 4096 matrix from each motif. These 50 coefficients were then regressed against the RS associated with each motif presented to that cell, using the following

general linear model: $r_m = a_1 w_1 + ... + a_n w_n + c$, where r_m is the response strength of a given cell to motif m, with m ranging from 1 to the total number of motifs presented to the cell, a_1 is the regression coefficient for the first wavelet coefficient, w_1, for coefficients 1 through n=50, and c is a constant term. For a given cell, the extent to which these 50 wavelet coefficients predict variation in the neural response can be assessed using R^2. As reported in the text, the mean R^2 for selective cells is significantly greater than that for non-selective cells. In addition, the regressions were significant for 3/16 (18.8%) non-selective cells and 17/28 (60.1%) song-selective cells, and the difference between these proportions is significant ($p < 0.01$, Chi-square). The mean R^2 for all cells (n = 44) was 0.63 ± 0.03. These R^2 values are not adjusted, and thus are likely inflated by the inclusion of so many coefficients in our regression model. We therefore also calculated R^2 values that were adjusted for the number of regressors and observations: $R^2_{adj} = 1-((1-R^2)(n-1)/(n-k-1))$, where n is the number of observations and k is the number of regressors. As expected, given the large differences in proportions of significant regressions between selective and non-selective cells and the fact that the same number of regressors were used in the model for each cell, the mean adjusted R^2 for the selective cells, while lower than for unadjusted R^2 values (0.297±0.045), was significantly greater than that for the non-selective cells (0.145±0.058; $p<0.05$, Mann-Whitney U-test).

The significant regression coefficients in our model indicate stimulus features represented by the wavelet coefficients that are correlated with neural responses. The mean number of significant regression coefficients for all cells was 4.96 ± 0.53. Because only an average of 5 out of 50 regressors were significant, and we chose the 50 strongest wavelet coefficients, it is unlikely that wavelet coefficients not included in our analysis contributed significantly to variation in motif RS. The one cell selective for synthetic stimuli was not included in any of the regression analyses. It is interesting to note the improvement in response predictions over the STRF method despite the fact

that both techniques rely on linear interpolation. The 2-D Haar transform and average response strength were calculated for entire motifs, whereas the time windows in the STRF method were limited to 200 ms preceding individual spikes.

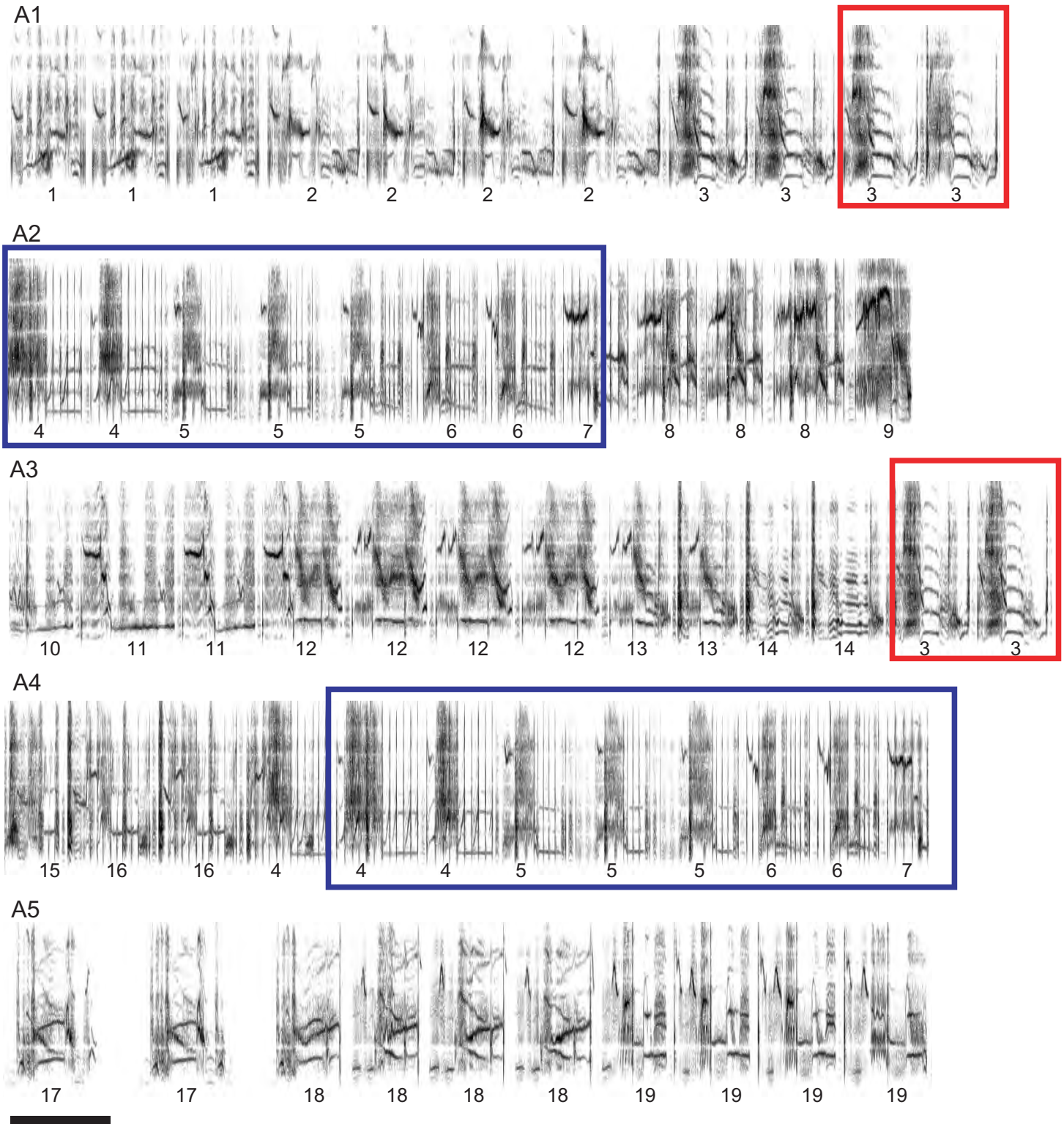
1. Eens, M., Pinxten, R., Verheyen, R. F. Temporal And Sequential Organization Of Song Bouts In The Starling. *Ardea* **77**, 75-86 (1989).
2. Adret-Hausberger, M. & Jenkins, P. F. Complex organization of the warbling song in the European starling, Sturnus vulgaris. *Behaviour* **107**, 138-156 (1988).
3. Lewicki, M. S. Bayesian modeling and classification of neural signals. *Neural Comp.* **6**, 1005-1030 (1994).
4. Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., Gallant, J.L. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network.* 12, 289 (2001).
5. Sen, K., Theunissen, F.E., Doupe, A. J. Feature analysis of natural sounds in the songbird auditory forebrain. *J Neurophysiol.* **86**, 1445 (2001).

Figure S1. Starling Song Stimuli. (a-c) Sonograms of the15 samples of male starling song used as stimuli in this study. Label letters denote the singer of each sample, and numbers denote the different exemplars from each set. Song labels correspond to those in Table S1. All the songs from a given singer comprise one set. Red and blue boxes show the regions of motif overlap within a set of songs, with matching colours for closely matching motifs. The same number under each motif indicates matching motifs, as we scored them. Scale bars show 1 s. Frequency range 0 –10 kHz.

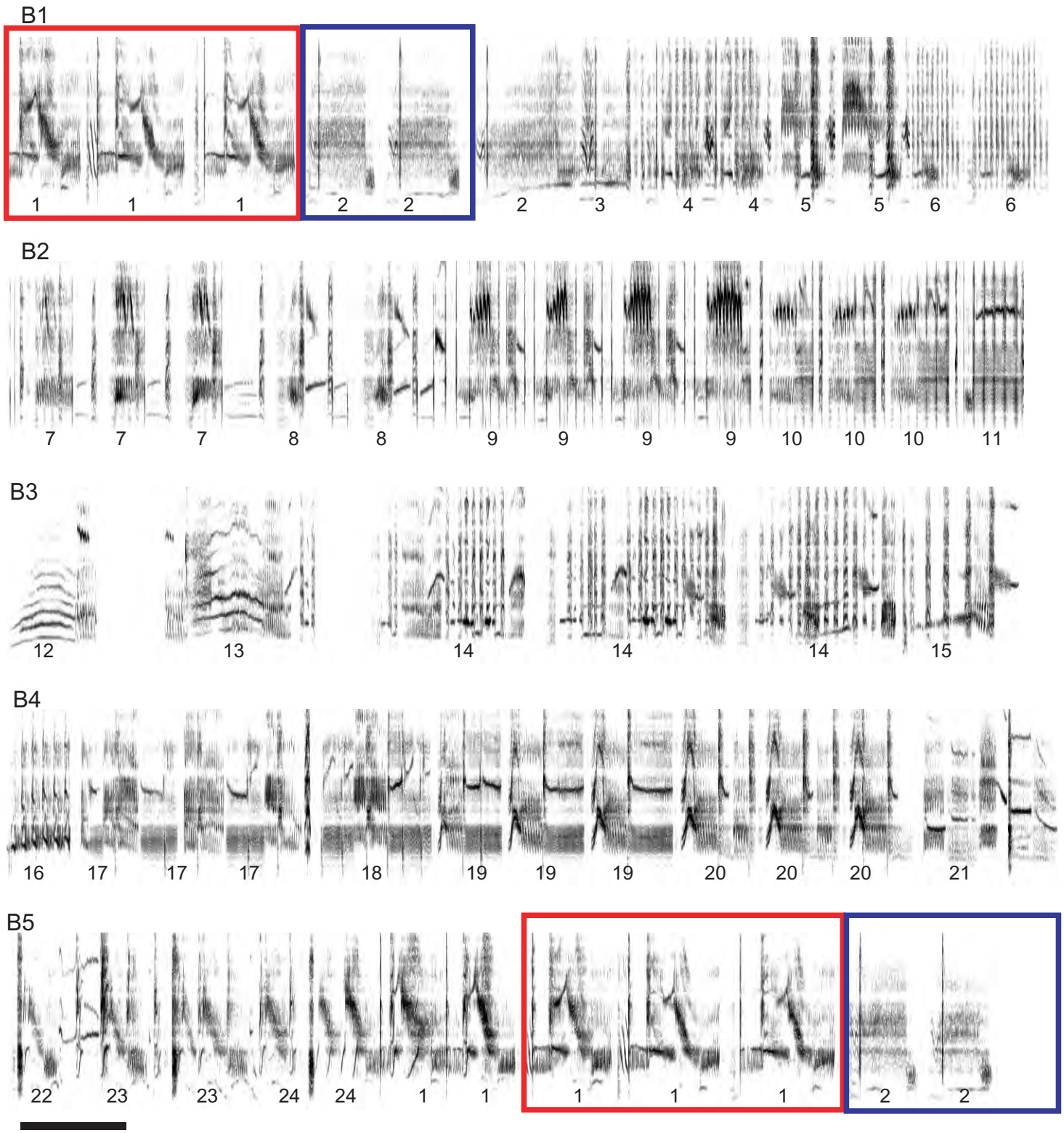
Figure S2. Spectrotemporal Receptive Fields. (a, b) Examples of two STRFs similar to those calculated for most of the cells in our sample, showing the optimal linear estimate of each cells preferred stimulus. Time is plotted in ms prior to the spike at time zero. Red shows the spectrotemporal regions of high excitation, and blue regions of inhibition. (c) STRF predicted (red) and actual (blue) firing rates plotted for the stimulus response depicted in Fig 2 A. Firing rates are shown as the deviation from the mean, and the PSTH of the actual rate is smoothed. (d) As above in (c) but for a cell and stimulus in which the cross-correlation is approximately equal to that for the entire data set.

Figure S3. Phasic response. Bivariate scattergram showing the strong negative correlation ($r = -0.76$, $p < 0.0001$, Fisher's r to z) between each cells phasic response (PR) and mean stimulus-driven spike rate. The line shows the linear regression.

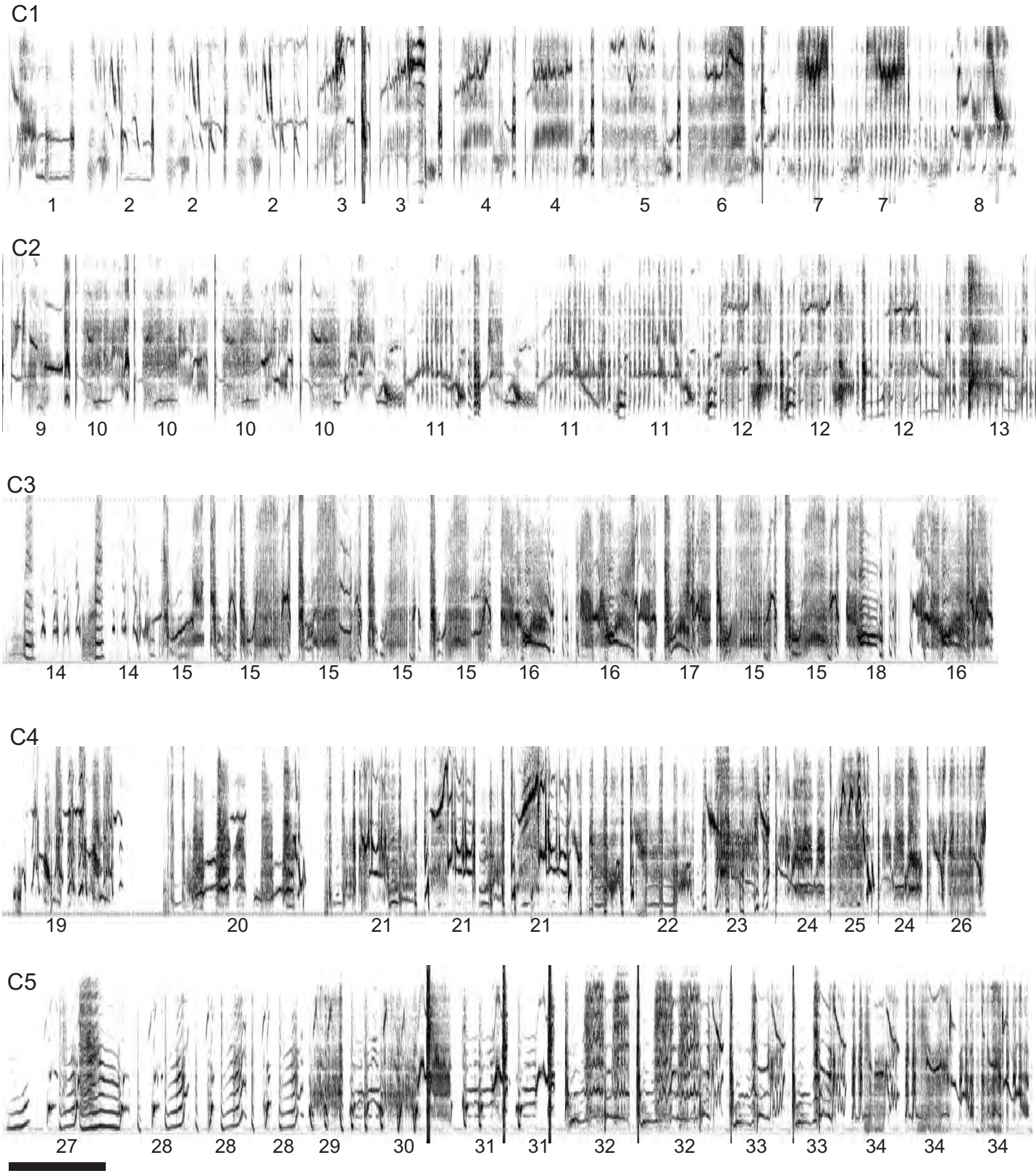
Gentner/2002-09188B SI Fig. 1, panel a

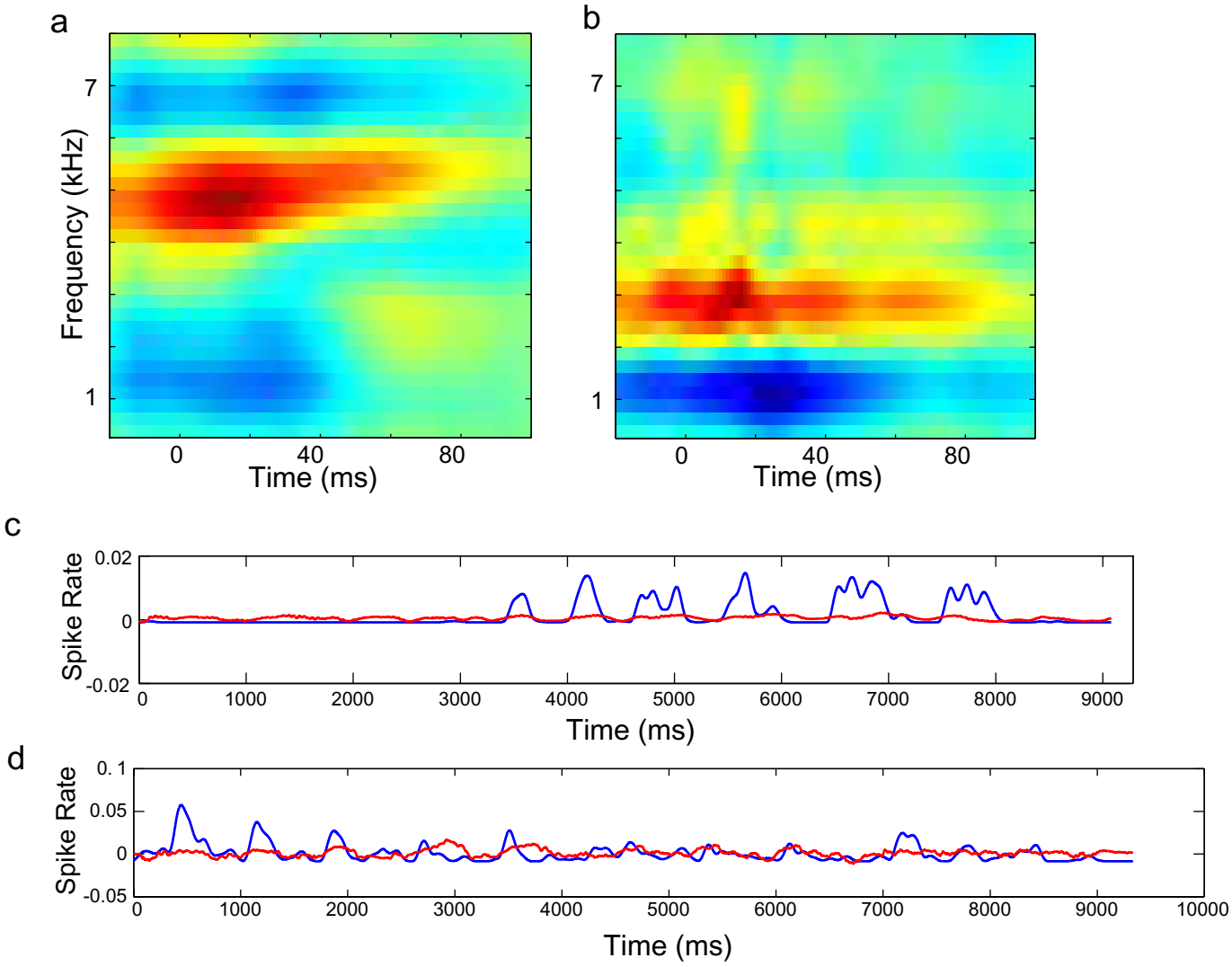


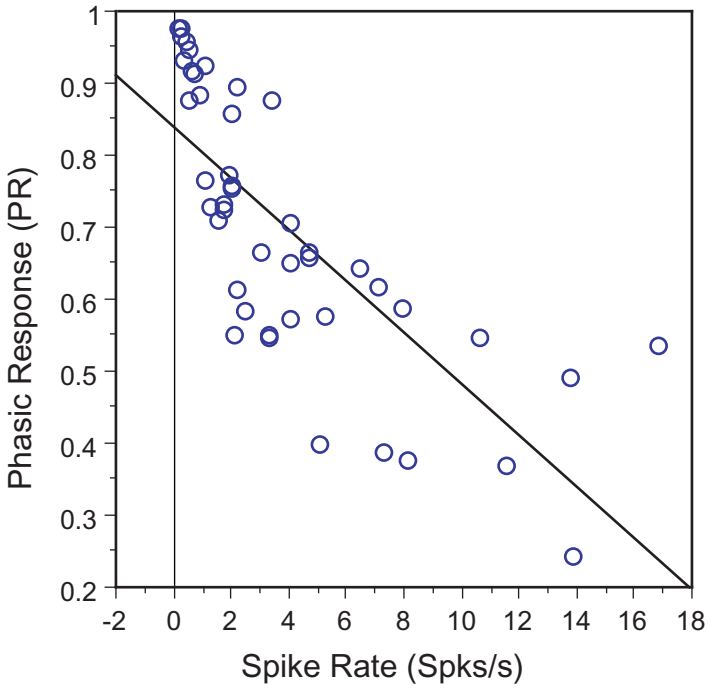
Gentner/2002-09188B SI Fig. 1, panel b



Gentner/2002-09188B SI Fig. 1, panel c




a
b
7
1
Frequency (kHz)
0
40
80
Time (ms)
c
Spike Rate
0.02
0
-0.02
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Time (ms)
d
Spike Rate
0. 1
0.05
0
-0.05
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Time (ms)


Phasic Response (PR)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
-2
0
2
4
6
8
10
12
14
16
18
Spike Rate (Spks/s)