Experimental and statistical reevaluation provides no evidence for *Drosophila* courtship song rhythms

David L. Stern[a,1], Jan Clemens[b], Philip Coen[c], Adam J. Calhoun[b], John B. Hogenesch[d], Ben J. Arthur[a], and Mala Murthy[b,e]

[a]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147; [b]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544; [c]University College London, Institute of Neurology, London WC1E 6BT, United Kingdom; [d]Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229; and [e]Department of Molecular Biology, Princeton University, Princeton, NJ 08544

Edited by Joseph S. Takahashi, Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, and approved August 9, 2017 (received for review May 6, 2017)

From 1980 to 1992, a series of influential papers reported on the discovery, genetics, and evolution of a periodic cycling of the interval between *Drosophila* male courtship song pulses. The molecular mechanisms underlying this periodicity were never described. To reinitiate investigation of this phenomenon, we previously performed automated segmentation of songs but failed to detect the proposed rhythm [Arthur BJ, et al. (2013) *BMC Biol* 11:11; Stern DL (2014) *BMC Biol* 12:38]. Kyriacou et al. [Kyriacou CP, et al. (2017) *Proc Natl Acad Sci USA* 114:1970–1975] report that we failed to detect song rhythms because (*i*) our flies did not sing enough and (*ii*) our segmenter did not identify many of the song pulses. Kyriacou et al. manually annotated a subset of our recordings and reported that two strains displayed rhythms with genotype-specific periodicity, in agreement with their original reports. We cannot replicate this finding and show that the manually annotated data, the original automatically segmented data, and a new dataset provide no evidence for either the existence of song rhythms or song periodicity differences between genotypes. Furthermore, we have reexamined our methods and analysis and find that our automated segmentation method was not biased to prevent detection of putative song periodicity. We conclude that there is no evidence for the existence of *Drosophila* courtship song rhythms.

Drosophila | courtship song | song rhythms

When a male vinegar fly (*Drosophila melanogaster*) encounters a sexually receptive female, he performs a series of courtship behaviors, including the production of songs containing pulses and hums (or sines) via unilateral wing vibration (Fig. 1*A*). Every parameter of song displays extensive quantitative variation within a bout of singing, including the amplitude and frequency of pulses and sines and the timing of individual pulse and sine events (1–8). Like humans during conversation, *Drosophila* males modulate their song based on sensory feedback from their communication partner (3, 4).

Visual inspection of songs reveals that the mean interpulse interval varies over time (Fig. 1*B*). This observation was first made in 1980 by Kyriacou and Hall (9) and they reported that the mean cycled with a periodicity of about 55 s and was controlled, in part, by the *period* gene, a gene required for circadian rhythms (10). Later papers demonstrated that evolution of a short amino acid sequence within the *period* protein caused species-specific differences in this periodicity (10–13). These reports attracted considerable interest because they implicated the *period* gene in ultradian rhythms, in addition to its well-known role in circadian rhythms (14), and because they illustrated how genetic evolution can cause behavioral evolution.

Despite this progress, the molecular mechanisms causing this periodicity remained unknown. To further advance study of these rhythms, previously we searched for this periodicity using sensitive methods and failed to find evidence for song rhythms (1). We were mindful, however, that Kyriacou and Hall (15) had argued that the presence or detectability of the rhythms was sensitive to assay conditions and methods of analysis. One of us, therefore, replicated the methods of Kyriacou and Hall as closely as possible, but, again, song rhythms could not be detected (2).

Kyriacou et al. (16) have recently questioned our previous conclusions. Here, we focus on three major assertions that they claim call our conclusions into doubt. First, we examine their central claim that manual analysis of songs, but not automated analysis, reveals genotype-specific song rhythms. We find that reanalysis of their manually annotated data provides no statistical support for genotype-specific rhythms. We also find no evidence for song rhythms in the original dataset and a new larger dataset. Second, we examined their claim that the original recordings contained insufficient data to detect rhythms and find that this claim is not supported by simulation studies. Third, we examine their claim that the high false-negative rate of the automated song segmenter decreased the probability of detecting song rhythms and we find no evidence that the missing pulse events biased our analysis of song rhythms. Further, we identify the major sources of false-negative events in automated song analysis and illustrate that minor modifications to initialization parameters substantially improve performance of the song segmenter. Kyriacou et al. (16) also raised a number of minor concerns—such as how to choose an appropriate interpulse-interval cutoff, whether temperature was controlled appropriately in our experiments, and whether songs produced beyond the first few minutes of courtship should be analyzed—that we consider peripheral to the central questions raised and therefore we have addressed these concerns (which are also unsupported by reanalysis) in *SI Appendix*.

Results

Earlier papers that identified song cycles used several unusual methods of data analysis that are useful to review. First, continuous

Significance

Previous studies have reported that male vinegar flies sing courtship songs with a periodic rhythm of approximately 55 s. Several years ago, we showed that we could not replicate this observation. Recently, the original authors have claimed that we failed to find rhythms because (*i*) our flies did not sing enough and (*ii*) our software for detecting song did not detect all song events. They reported that they could detect rhythms in song annotated by hand. We show here that we cannot replicate their observation of rhythms in the hand-annotated data or in other datasets. We also show that our original methods were not biased against detecting rhythms. We conclude that song rhythms cannot be detected.

Author contributions: D.L.S., J.C., P.C., A.J.C., J.B.H., B.J.A., and M.M. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The raw and segmented song data for the new song recordings are available in the Dryad Digital Repository, datadryad.org/ (doi:10.5061/dryad.td223).

[1]To whom correspondence should be addressed. Email: sternd@janelia.hhmi.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1707471114/-/DCSupplemental.

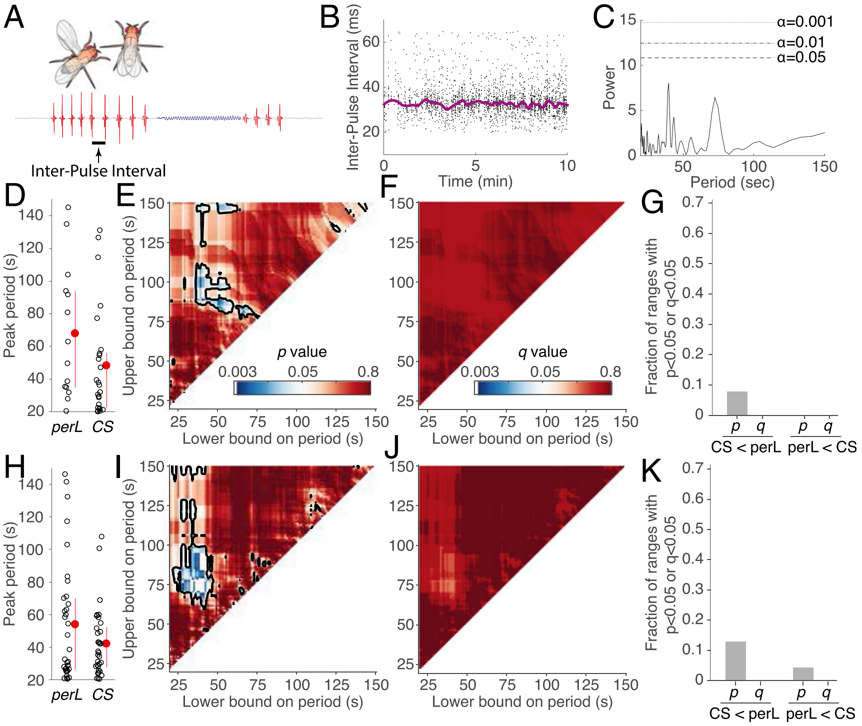


Fig. 1. Genotype-specific periodicity cannot be detected in *Drosophila* courtship song. (*A*) *Drosophila* males produce courtship song, composed of pulses (red) and sines (blue), by extending and vibrating a wing. The interpulse interval is the time between consecutive pulses within a single train of pulses. (*B*) The average interpulse interval varies over time. (Purple line is the running mean with sliding window of 200 samples.) (*C*) Lomb–Scargle periodogram analysis of the interpulse-interval data from *B* plotted for the range of 20–150 s. None of the peaks are significant at $P < 0.05$. (*D*) Comparison of the peak power between 20 and 150 s from the Lomb–Scargle periodograms for the song data for the genotypes periodL (perL) and Canton-S (CS) manually annotated by Kyriacou et al. (16). Red points and lines represent mean ± 1 SD for each genotype. (Right-tailed *t* test $P = 0.06$. Rank sum $P = 0.10$.) (*E*) *P* values for period windows with different lower and upper bounds. (*F*) False discovery rate *q* values for the windows shown in *E*. (*G*) Fraction of ranges with significant comparisons (*p* or $q < 0.05$) for either the test of Canton-S less than periodL or periodL less than Canton-S. (*H*–*K*) Same as *D*–*G* for newly collected song data from the same genotypes annotated using FlySongSegmenter. (*H*) Right-tailed *t* test $P = 0.06$; rank sum $P = 0.45$.

interpulse-interval data were binned into 10-s intervals. We reported previously that binning the data, together with the analysis of relatively short songs, creates peaks in spectrogram analysis that fall within an artificially narrowed frequency range, corresponding approximately to the frequency range originally reported for the periodicity, and reduces the significance of periodiogram peaks (ref. 2 and see below). Despite the fact that this procedure squeezes periodogram results into a narrow frequency range, few songs contained peaks reaching a significance level of $P < 0.05$ (4 of 149 songs, figure 3*A* of ref. 2), strongly suggesting that these peaks represent signals that cannot be distinguished from noise. All of the previously reported "statistically significant" comparisons of different genotypes are derived from analysis of mainly nonsignificant periodogram peaks. In this reevaluation, we do not discuss binning, but instead focus on other methodological issues.

No Evidence That Manual Song Segmentation Reveals Genotype-Specific Song Rhythms. Kyriacou et al.'s (16) core finding is that different genotypes displayed different periodic rhythms of the interpulse interval. This is also the most important discovery reported in earlier papers on this subject (10–12, 17). Kyriacou et al. (16) manually annotated recordings made by Stern (2) from a wild-type strain, *Canton-S*, and a strain carrying a *period* gene mutation, *per*^{L}, for flies they categorized as singing "vigorously." We reanalyzed these data and the automatically segmented data (2). Flies homozygous for *per*^{L} display circadian rhythms that are longer than normal (14), and earlier papers have reported that *per*^{L} confers longer periods on the interpulse-interval rhythm (9–12). Kyriacou et al. (16) report a difference in the mean song period between *Canton*-S and *per*^{L} with the manually annotated data, but not with the automatically segmented data, suggesting that song cycles exist and display genotype-specific frequencies and that the automatically segmented data are biased against detecting the song rhythm.

Kyriacou et al. (16) used several methods to measure periodicity in the original time series, which we discuss in more detail in the next paragraph. For ≈85% of these songs, these methods do not yield statistically significant signals in the frequency range of 20–150 s. Because most songs do not yield statistically significant peaks, Kyriacou et al. (16) identified the peak with maximum power in the range of 20–150 s for each song and compared these values between genotypes. This is an unorthodox approach to data analysis. It is equivalent to sampling outliers from a distribution of random noise and then performing further statistics with these data. Nonetheless, Kyriacou et al. (16) detected genotype-specific song rhythms using this method and so, below, we accept this premise and investigate whether there is statistical support for genotype-specific rhythms in the data. We start by examining whether there is evidence for rhythms in individual songs.

The general model proposed for these song rhythms is that the interpulse interval varies, on average, with a regular periodicity (9). Therefore, it should be possible to detect this rhythmicity with appropriate methods of periodogram analysis. We have previously used Lomb–Scargle periodogram analysis (18–20) because this method does not require evenly spaced samples and Kyriacou et al. (16) also adopted this method. For example, the Lomb–Scargle periodogram of the time series in Fig. 1*B* is shown in Fig. 1*C*. In this case, despite the obvious variation in interpulse-interval values observed in Fig. 1*B*, there is no significant periodicity between 20 and

150 s. Kyriacou et al. (16) also used Cosinor (21) and CLEAN (22) for periodogram analysis. CLEAN does not produce a significance value for periodogram peaks, so it is difficult to interpret. We find that Cosinor exhibits a high false-positive rate (*SI Appendix*, Fig. S1) and should be avoided for this type of analysis.

Kyriacou et al. (16) state that wild-type *D. melanogaster* songs exhibit periodicity between 20 and 150 s. Previously, they reported that rhythms occurred with 50–60 s periodicity (9). Increasing the width of the periodicity window from 50–60 s to 20–150 s increases the probability of detecting significant periods, but, even given this wide frequency range, we observed that only 4 of the 25 manually annotated *Canton-S* songs and 3 of the 25 automatically segmented songs contained periodogram peaks that reached a significance level of $P < 0.05$. (When we binned data in 10-s bins, these values declined to 0 of 25 manually annotated and 1 of 25 automatically segmented songs.) These significant peaks are not localized to any particular narrow frequency range (*SI Appendix*, Figs. S1 and S5).

One reason to study nonsignificant peaks would be if periodicity is weak and not detected reliably by periodogram analysis. This seems unlikely, since simulated song rhythms can be detected with high confidence (refs. 1 and 2 and see below). Nonetheless, if periodogram analysis is underpowered, then we expect to observe that the major peak in most songs should display nearly significant periodicity. In fact, we observe that 72% of P values are greater than 0.2 (*SI Appendix*, Fig. S2). There is therefore no evidence that songs contain weak periodicity.

An alternative possible reason to include nonsignificant periodogram peaks in downstream analysis is that the signal to noise of the periodicity is extremely low. An analog in neuroscience is that neural signals sometimes cannot be detected with high signal to noise and that only by averaging over many trials of a stimulus presentation can a neural response be detected robustly. We therefore examined the power distribution averaged over all of the results for each genotype. These plots are essentially flat, suggesting that there is no signal hidden in the fluctuations of individual periodograms (*SI Appendix*, Fig. S3).

Given these observations, further analysis of these data seems unwarranted. However, Kyriacou et al. (16) compared the maximum periodogram peaks between 20 and 150 s for the *Canton-S* and *per^L* recordings and found that the manually annotated data showed a statistically significant difference in the mean period, although the automatically segmented data did not (figure 3*D* of Kyriacou et al.; ref. 16). This is the key result of their paper. We therefore attempted to replicate this observation. For the manually annotated data from each song, we identified the peak in the periodogram of maximum power falling between a period of 20 and 150 s. In contrast to their published results, we found that the average of the periods with maximum power (most of which were not significant) was not significantly different at $P < 0.05$ between the genotypes *Canton-S* and *per^L* (Fig. 1*D*). We have no explanation for this discrepancy between our statistical analysis and theirs.

Because there is no biological or quantitative justification for the particular frequency ranges examined in any study, we wondered whether the results were sensitive to the frequency range examined. We explored a wide range of possible frequency ranges and found that the test statistic was sensitive to the precise frequency range selected (Fig. 1*E*). Most frequency windows do not generate a statistically significant difference between the genotypes (Fig. 1 *E* and *G*), and false discovery rate correction for multiple testing (23, 24) yields no frequency ranges with significant results (Fig. 1 *F* and *G*).

Thus, there is no support for the specific results reported by Kyriacou et al. (16) and there is no statistical support for defining song interpulse-interval cycle periods as occurring within any particular window. Most importantly, our analysis indicates that genotype-specific analysis of nonsignificant periodogram peaks has no justification. It is difficult to reconstruct precisely what steps in the analysis led previous reports to identify statistically significant genotype-specific differences, but it is possible that previous studies may have serendipitously selected frequency ranges that yielded significant results and/or did not properly control for multiple testing.

Newly Collected Data Provide No Evidence for Genotype Specific Song Periodicities. Although we could not reproduce results reported by Kyriacou et al. (16), we decided to take their observation at face value as a preliminary result and test directly whether genotype specific song rhythms could be detected in an expanded dataset. We recorded song from 33 *Canton-S* males and 34 *period^L* males. We identified the strongest periodogram peak in the frequency range of 20–150 s for each song and found no significant difference between these genotypes (Fig. 1*H*). We then compared test statistics across a wide set of frequency ranges, as described above. We identified some frequency ranges that yielded significant results in the predicted direction (Fig. 1*I*), with *period^L* rhythms slower than *Canton-S* rhythms, but for three reasons we believe these results are spurious. First, and most importantly, none of these ranges are significant after false discovery rate correction (Fig. 1*J*). Second, multiple frequency ranges support the opposite conclusion, that *Canton-S* rhythms are slower than *period^L* rhythms (Fig. 1*K*). Third, the frequency ranges yielding significant comparisons only partially overlap with the ranges found for the original dataset (cf. Fig. 1 *E* and *I*). In conclusion, there is not only no evidence that song rhythms exist, there is also no evidence that reported genotype-specific differences in a song rhythm exist.

Putative song cycles cannot be identified in most automatically segmented song (2) and, as we showed above, in most manually annotated song. In addition, when statistically significant periodicity is detected, the frequencies of this periodicity do not cluster in a specific frequency range, but instead are spread randomly across the entire frequency range examined (*SI Appendix*, Fig. S5; figure 4 of Stern; ref. 2). Finally, no genotype comparisons are significant after correcting for multiple comparisons (Fig. 1). All together, these results imply that the few statistically significant periodicities that can be found do not carry biological significance.

No Evidence That Low-Intensity Courtship Provided Insufficient Data to Detect Song Rhythms. Although we found no statistical evidence for the existence of song rhythms or of genotype-specific rhythms, we feel it is important to rebut several other statements made by Kyriacou et al. (16). They state that rhythms can be detected only in songs produced by vigorously singing males and write: "sporadic songs could not possibly provide any test for song cycles." It is not clear if they mean that rhythms can be detected only in songs with many pulses or that only flies that sing songs with many pulses ("vigorous singers") produce rhythms. Kyriacou et al. (16) manually annotated songs from flies that they categorized as vigorous, and we showed above that significant periodicity can be found in only a minority of these songs and that these significant values are not localized to a particular frequency range (*SI Appendix*, Figs. S1*D* and S5*A*). Therefore, it is unlikely that only flies that sing songs with many pulses produce periodicity. We therefore performed simulations to determine whether rhythms can be detected only in songs with many pulses.

We previously investigated songs from 45-min courtship recordings that contained at least 1,000 interpulse-interval measurements (2). Kyriacou et al. (16) argued that more than 180 interpulse-interval measurements per minute (or ≈5,000 events in a 45-min recording) should be identified to allow identification of song rhythms. To examine this claim, we performed a statistical power analysis using songs with variable numbers of interpulse-interval measurements, where statistical power corresponds to the proportion of times periodicity is detected in songs where periodicity has been artificially imposed on song data (Fig. 2). We started with six 45-min recordings of *Canton-S* from Stern (2) that contained more than 10,000 interpulse-interval measurements. None of these six songs yielded statistically significant power in the frequency range between 50 and 60 s (the range originally defined to contain rhythms; ref. 9) and one song produced a marginally significant peak at 31.7 s ($P = 0.04$), which falls between 20 and 150 s (the range used by Kyriacou et al.; ref. 16). Fig. 2 *D* and *E* illustrate the interpulse-interval data and periodogram for one of these songs. Therefore, these songs do not contain strong periodicity in the predicted range and can serve as a template to examine the power of Lomb–Scargle periodogram analysis to detect simulated rhythms imposed on these data.

The initial reports of periodic cycles in the interpulse-interval reported rhythms with a mean period of 55 s and an amplitude of ≈2 ms (9). Therefore, we imposed a 55-s rhythm with an amplitude of 2 ms on the six songs containing more than 10,000 interpulse-interval measurements (Fig. 2 *A*–*C*). We detected the simulated 55-s rhythm in all six songs with *P* values <10*e*-74 (example shown in Fig. 2 *F* and *G*). We then randomly removed data points from the songs iteratively and calculated the fraction of times we could detect the simulated rhythm with $P < 0.05$. We removed data randomly from the dataset to simulate the effect of failing to detect individual events in the song, and we also removed chunks of data (in 10-s bins) to simulate large gaps between song bursts, such as might be generated during low-intensity courtship. We found that in both scenarios we could randomly remove at least 90% of the data and still detect simulated rhythms at least 80% of the time (example shown in Fig. 2 *H* and *I*; summary statistics shown in Fig. 2*J* and *SI Appendix*, Fig. S4*A*). That is, as long as songs contained at least 1,000 interpulse-interval measurements, Lomb–Scargle periodogram analysis detected simulated rhythms with power greater than 0.8. Similar results were found when we analyzed only the first 400 s of songs (*SI Appendix*, Fig. S4 *C* and *D*). Furthermore, periodicity could be detected with power greater than 0.8 when the amplitude of simulated periodicity was greater than at least 1 ms (*SI Appendix*, Fig. 4*B*). These results were robust to noise in the original periodicity. Song with a signal-to-noise ratio of as low as 0.25 could be detected with power >0.7 with sample sizes of at least 1,000 interpulse-interval measurements (Fig. 2*K*). Similarly, periodicity could be detected reliably when we simulated a nonsinusoidal rhythm (*SI Appendix*, Fig. S4*E*) and when periodicity was imposed for only a fraction of the total song (*SI Appendix*). Thus, Lomb–Scargle periodogram analysis is a sensitive method for detecting simulated periodicity, even in the presence of noise or discontinuities in the waveform.

Songs containing at least 1,000 interpulse intervals provide sufficient data to identify putative song cycles. In fact, we find that songs can be deeply corrupted by the absence of large segments of song and simulated periodicity can still be detected.

No Evidence That the Automated Fly Song Segmenter Biased the Results. Kyriacou et al. (16) expressed concern that our automated fly song segmenter displayed a low true positive rate (the segmenter failed to detect ≈50% of the pulses identified through manual annotation) and produced some false-positive calls (≈4% of events scored as pulses by the automated segmenter appear to be noise). They suggest that these incorrect pulse event assignments could bias estimation of the mean interpulse interval and, therefore, decrease the signal to noise of the periodic cycle, making it difficult to detect a periodic signal. In principle, a large

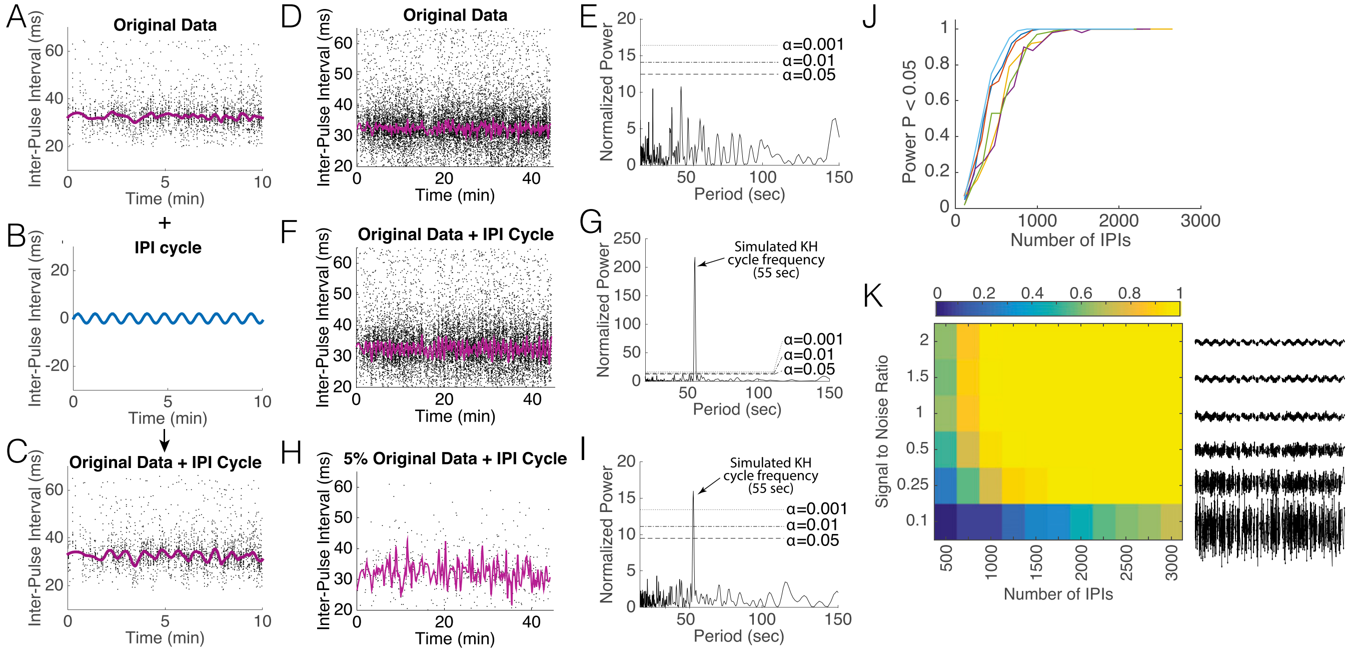


Fig. 2. Simulations to explore power to detect rhythms, should they exist. (*A*–*C*) Example of how a periodic cycle was added to raw interpulse-interval (IPI) data. Purple line in *A* illustrates the running mean of the raw data. Blue line in *B* shows a periodic rhythm with an amplitude of 2 ms and a period of 55 s. Original data with simulated periodicity is shown in *C*. (*D*) One example of 45 min of interpulse-interval data. Purple line shows running mean. (*E*) Lomb–Scargle periodogram of data in *D* does not detect periodicity. (*F*) Data from *D* with a 55-s periodicity imposed. (*G*) Lomb–Scargle periodogram of data in *F* now reveals a highly significant peak at 55 s, consistent with the simulated Kyriacou–Hall (KH) periodicity. (*H*) Random removal of 95% of the interpulse-interval data from *F*. (*I*) Lomb–Scargle periodogram of the data in *H* detects significant periodicity. (*J*) Power analysis of six songs (each song a different color) containing more than 10,000 interpulse-interval events after 55-s periodicity was added and individual interpulse-interval events were removed randomly. Power equals the fraction of times out of 100 that a song contained a rhythm with significant periodicity between 50 and 60 s at $P < 0.05$. (*K*) Power to detect simulated noisy periodicity versus number of IPIs remaining after random removal of IPIs. Means of simulations for six songs containing more than 10,000 interpulse-interval measurements are shown. Examples of simulated noisy rhythms are shown to the *Right*. Colorbar shows power to detect simulated rhythm.

sample of incorrect calls could bias results, so we investigated whether this was the case for our prior analyses. We used Kyriacou et al.'s (16) manually annotated dataset to investigate the potential for bias and to evaluate performance of the automated segmenter.

When a single pulse event is not detected, the interpulse interval is then calculated as the sum of the two neighboring real intervals. On average, this is approximately double the average interpulse interval. The average interpulse interval for the *Canton-S* recordings reported in Stern (2) is ≈35 ms with a SD of ≈7 ms. Therefore, skipping a single pulse event is expected to result in interpulse-interval measurements of ≈70 ms, but with considerable variance. Following Kyriacou and Hall (15) and Stern (2), we used a heuristic threshold of 65 ms to reduce the number of spurious interpulse-interval values. Therefore, in the specific case when a single pulse in a train is missed, approximately one-third of the incorrectly scored doublet interpulse-interval measurements would be shorter than 65 ms and are expected to contaminate the original dataset.

However, this scenario applies only when one undetected pulse is flanked by two pulses that are detected. Skipping more than one pulse would always result in interpulse-interval measurements that are excluded by the 65-ms threshold. We found, however, that only 9% of the pulses missed by automated segmentation were singletons (*SI Appendix*, Fig. S6*A*). These incorrect interpulse intervals contribute to a slight excess of interpulse intervals with high values (*SI Appendix*, Fig. S6*B*). Lowering the interpulse-interval threshold would, therefore, remove most or all spurious interpulse intervals. Since our power analysis, discussed above, revealed that periodogram analysis was robust to random removal of interpulse-interval events, as long as songs still contained at least 1,000 values, loss of a small number of interpulse intervals is not expected to hamper detection of rhythms. After reducing the interpulse-interval threshold to 55 ms, we still found no compelling evidence for significant periodicity in the original data (*SI Appendix*, Fig. S7). Therefore, we explored the effect of reducing the interpulse-interval cutoff even further. In this case, we used all 68 *Canton-S* songs from Stern (2) and retained for analysis only those songs that contained at least 1,000 interpulse-interval measurements after imposing the new interpulse-interval threshold. We explored a range of cutoff values from 25 to 65 ms. We found that we could detect the simulated rhythm in most songs with at least 1,000 interpulse-interval measurements remaining after thresholding, even when the threshold was as low as 25 ms (Fig. 3). Therefore, we can find no evidence that pulses missed by the automated song segmenter or the specific interpulse-interval threshold used in Stern (2) prevented detection of song rhythms.

Although detection of putative song rhythms is robust to dropped pulses in songs that retain at least ≈1,000 interpulse intervals, it is worth reviewing briefly why the segmenter failed to detect certain pulses in recordings reported in Stern (2). The first step of song segmentation involves detection of pulse-like signals and sine-like signals (1). In subsequent steps, the segmenter filters out many kinds of sounds that were originally classified as song pulses. Both the initial detection of pulses and subsequent filtering steps are sensitive to multiple parameters. These parameters are specified before segmentation and can be modified to enhance performance of the segmenter for different recordings. We identified two primary causes for missed pulses. First, Stern (2) recorded song in larger chambers than those used previously with these microphones (1), to match the chamber size used by Kyriacou and Hall (9). This larger chamber with one microphone had reduced sensitivity compared with the original smaller chamber. The segmenter thus tended to miss pulses of lower amplitude, which are hard to automatically differentiate from noise, and this explains ≈35% of the missed pulses (*SI Appendix*, Fig. S8 *A* and *C*).

The second major cause of missed pulses is that *Drosophila* males produce pulses with a range of carrier frequencies (tones). The higher frequency pulses tend to resemble other nonsong noises, like grooming, and a user can set parameters in the segmenter to attempt to exclude these nonsong noises based on the carrier frequency of the event. Stern (2) used parameters to minimize the false-positive rate, including a relatively low carrier frequency cutoff for pulses. The lower pulse frequency threshold used by Stern (2) explains ≈42% of the missed pulses (*SI Appendix*, Fig. S8 *B* and *D*). Using the same software with different parameters (from Coen et al.; ref. 4) recovers many of these high-frequency pulses without substantially increasing the false-positive rate (*SI Appendix*, Fig. S8 *C*–*F*).

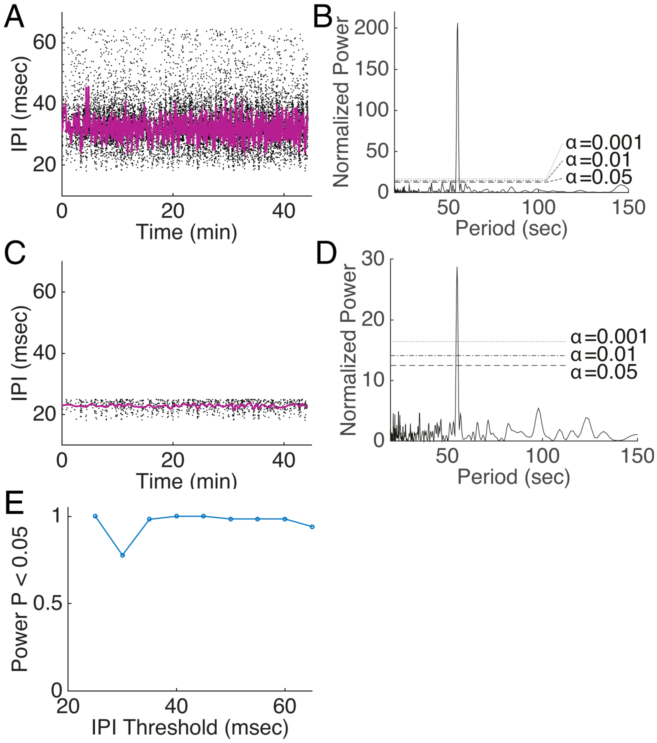


Fig. 3. The specific interpulse-interval threshold does not influence the statistical power to detect putative song rhythms. (*A*) Example of one original song with 55-s periodicity artificially imposed on the original interpulse-interval data. (*B*) Lomb–Scargle periodogram of data in *A*, revealing strong signal at 55 s. (*C*) Same simulated data as in *A* with all interpulse-interval values greater than 25 s removed. (*D*) Lomb–Scargle periodogram reveals strong signal of the simulated periodicity at 55 s, even though the data were thresholded at 25 s. (*E*) Power to detect simulated periodicity versus interpulse-interval threshold for songs retaining at least 1,000 interpulse-interval values after thresholding.

Above, we showed that including more pulse events, by manual annotation, did not increase the probability of detecting song rhythms. Therefore, there is no evidence that the data resulting from the song segmenter parameters used in Stern (2) generated a dataset that was biased against detection of song rhythms. While the song segmenter does not detect all pulse events that can be detected by manual annotation, the segmenter does provide datasets that are several orders of magnitude larger than those that can be generated by manual annotation, which has allowed discovery of multiple new phenomena related to *Drosophila* courtship song (3–7). In addition, the sensitivity of the song segmenter can be improved with optimization of initial parameters, as expected of any segmentation algorithm.

Discussion

We cannot detect a periodic cycling of the interpulse interval in *Drosophila* courtship song even in the songs manually annotated by Kyriacou et al. (16) and used as evidence for periodicity in their paper. Although it is impossible to prove a negative, our results agree with previous analyses that have concluded that there is no statistical evidence that these rhythms exist (1, 2). In particular, by exploring some of the relevant parameter space with statistical tests on the song that was manually annotated by Kyriacou et al. (16), we find that subsets of parameters sometimes produce P values lower than 0.05, but that (*i*) few regions of parameter space generate "significant" results, (*ii*) these significant regions are scattered apparently randomly in parameter space, and (*iii*) none of these significant results survive multiple test correction (Fig. 1).

Previously, we offered one explanation for how apparent song rhythms may have been detected. We found that binning data from short songs confined the periodogram peaks with maximum power close to the range reported as the song cycle (2). While few of these peaks reached statistical significance, previous authors have accepted these peaks as "signal" and performed statistical analyses to compare the peaks between genotypes. All statistically significant results from earlier papers were derived mainly from nonsignificant peaks in periodogram analysis and from relatively small sample sizes (usually fewer than 10 flies of each genotype), so it is questionable whether these derivative statistics are valid. Genotype-specific periodicities reported in earlier papers may have resulted, by chance, from studies of a small number of short songs that fortuitously led to occasional apparent replication of the original observations.

There may be a more prosaic explanation for the initial discovery of song cycles. Every fly produces highly variable interpulse intervals. In addition, a running average of these data reveals that the average interpulse-interval cycles up and down (Fig. 1*B*), similar to the temporally binned data first reported by Kyriacou and Hall (9). There is no debate about this observation. The claim in dispute is that the average interpulse-interval cycles regularly. We can find no evidence for this claim. It is easy to imagine, however, that visual examination of short recordings of song would make it appear as if the mean interpulse-interval cycled regularly.

The extraordinary within-fly variation in the interpulse interval and in the mean interpulse interval may result from multiple causes, including the possibility that male flies respond to ever-changing cues during courtship and modulate their interpulse interval to optimize their chances of mating. Individual *Drosophila* males modulate specific aspects of their courtship song based on their own patterns of locomotion and in response to feedback from females, including the transition between sine and pulse song (4) and the amplitude of pulse song (3). There is additional evidence that males modulate the carrier frequency of sine song (1). We hypothesize that male flies also modulate their interpulse interval in response to specific internal or external cues.

We can find no statistical evidence for periodicity of the interpulse interval in individual courtship songs and no evidence that comparisons of the strongest periodogram peaks from each song identify genotype-specific rhythms. These results hold both for the songs manually annotated by Kyriacou et al. (16) and for two independent large datasets automatically annotated with FlySongSegmenter using optimized parameters. At this time, a conservative assessment of the problem is that *Drosophila* courtship song rhythms and genotype-specific effects on these rhythms cannot be replicated.

Methods

Computer code for all analyses described in this paper is available at https://github.com/murthylab/noIPIcycles. Code for the version of FlySongSegmenter used in Cohen et al. (5) is available at https://github.com/murthylab/songSegmenter. The raw and segmented song data for the new song recordings are available at https://www.janelia.org/lab/stern-lab/tools-reagents-data. Further methods can be found in *SI Appendix*.

ACKNOWLEDGMENTS. We thank Elizabeth Kim for recording the new samples of flies. M.M. is a Howard Hughes Medical Institute Faculty Scholar and was also funded by an NIH New Innovator Award and an NSF Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative Early Concept Grant for Exploratory Research (EAGER) Award.

1. Arthur BJ, Sunayama-Morita T, Coen P, Murthy M, Stern DL (2013) Multi-channel acoustic recording and automated analysis of Drosophila courtship songs. *BMC Biol* 11:11.
2. Stern DL (2014) Reported Drosophila courtship song rhythms are artifacts of data analysis. *BMC Biol* 12:38.
3. Coen P, Xie M, Clemens J, Murthy M (2016) Sensorimotor transformations underlying variability in song intensity during Drosophila courtship. *Neuron* 89:629–644.
4. Coen P, et al. (2014) Dynamic sensory cues shape song structure in Drosophila. *Nature* 507:233–237.
5. Ding Y, Berrocal A, Morita T, Longden KD, Stern DL (2016) Natural courtship song variation caused by an intronic retroelement in an ion channel gene. *Nature* 536: 329–332.
6. Shirangi TR, Wong AM, Truman JW, Stern DL (2016) Doublesex regulates the connectivity of a neural circuit controlling Drosophila male courtship song. *Dev Cell* 37: 533–544.
7. Shirangi TR, Stern DL, Truman JW (2013) Motor control of Drosophila courtship song. *Cell Rep* 5:678–686.
8. Ewing AW, Bennet-Clark HC (1968) The courtship songs of Drosophila. *Behaviour* 31: 288–301.
9. Kyriacou CP, Hall JC (1980) Circadian rhythm mutations in Drosophila melanogaster affect short-term fluctuations in the male's courtship song. *Proc Natl Acad Sci USA* 77: 6729–6733.
10. Zehring WA, et al. (1984) P-element transformation with period locus DNA restores rhythmicity to mutant, arrhythmic Drosophila melanogaster. *Cell* 39:369–376.
11. Wheeler DA, et al. (1991) Molecular transfer of a species-specific behavior from Drosophila simulans to Drosophila melanogaster. *Science* 251:1082–1085.
12. Kyriacou CP, Hall JC (1986) Interspecific genetic control of courtship song production and reception in Drosophila. *Science* 232:494–497.
13. Ritchie MG, Halsey EJ, Gleason JM (1999) Drosophila song as a species-specific mating signal and the behavioural importance of Kyriacou & Hall cycles in D. melanogaster song. *Anim Behav* 58:649–657.
14. Konopka RJ, Benzer S (1971) Clock mutants of Drosophila melanogaster. *Proc Natl Acad Sci USA* 68:2112–2116.
15. Kyriacou CP, Hall JC (1989) Spectral analysis of Drosophila courtship song rhythms. *Anim Behav* 37:850–859.
16. Kyriacou CP, Green EW, Piffer A, Dowse HB (2017) Failure to reproduce period-dependent song cycles in Drosophila is due to poor automated pulse-detection and low-intensity courtship. *Proc Natl Acad Sci USA* 114:1970–1975.
17. Kyriacou CP, van den Berg MJ, Hall JC (1990) Drosophila courtship song cycles in normal and period mutant males revisited. *Behav Genet* 20:617–644.
18. Lomb NR (1976) Least-squares frequency analysis of unequally spaced data. *Astrophys Space Sci* 39:447–462.
19. Scargle JD (1982) Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys J* 263:835–853.
20. Ruf T (1999) The Lomb-Scargle periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series. *Biol Rhythm Res* 30:178–201.
21. Refinetti R, Lissen GC, Halberg F (2013) Procedures for numerical analysis of circadian rhythms. *Biol Rhythm Res* 38:275–325.
22. Roberts DH, Lehár J, Dreher JW (1987) Time series analysis with CLEAN. I. Derivation of a spectrum. *Astron J* 93:968–989.
23. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57:289–300.
24. Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* 1:140216.

Supplementary Information Appendix

The first section of this supplementary information appendix contains supplementary figures that are cited in the main paper. In addition, at the end of this appendix, we address several issues raised in Kyriacou et al. (1) that we did not have space to address in the main manuscript.

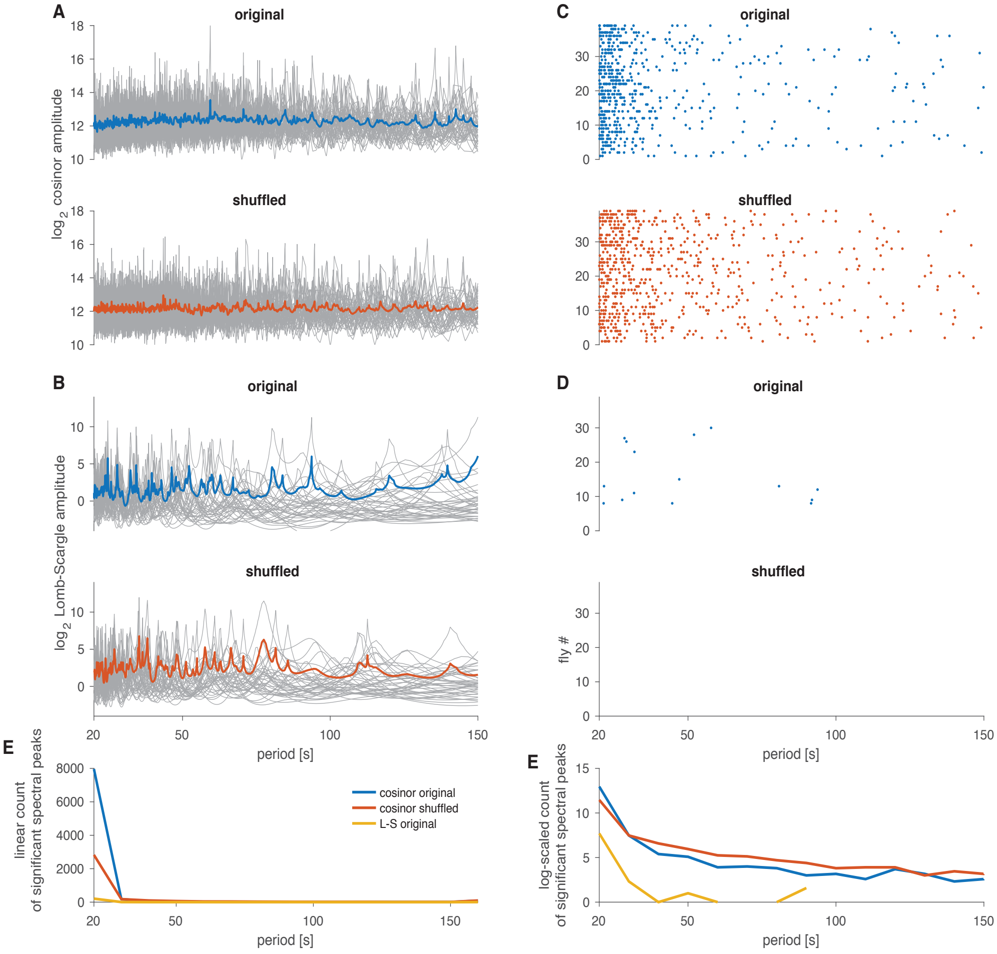


Figure S1. Cosinor analysis of IPI cycles produces many false positives. (A,B) Amplitude of cosinor fits (A) and Lomb-Scargle spectral power (B) for periods in the range 20 to 150 seconds (log2 scale). Spectra for original IPI data (upper panel) and shuffled IPI data (lower panel) are similar. Grey lines show spectra for individual flies, blue/orange lines show population averages for original and shuffled data, respectively. (C,D) Frequencies of significant peaks in the cosinor (C) and Lomb-Scargle (D) spectra for original (upper panel, blue dots) and shuffled data (lower panel, orange dots). One line per fly. (E) Distribution of significant periods over all flies shows an enrichment of short periods (20-30ms). (F) Same distribution as in E but with logarithmic y-scale to highlight counts for high periods. There is no enrichment for longer periods, suggesting that they are false positives.

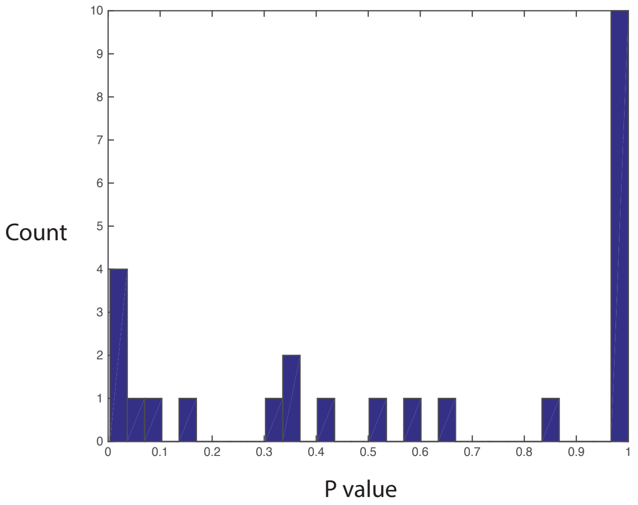


Figure S2. Distribution of p-values for the Lomb-Scargle periodogram peaks with maximum power between 20 and 150 sec for the *Canton-S* song data manually annotated by Kyriacou et al (3). Four of the peaks exhibit p-values < 0.05 and there is not an obvious excess of low p-values.

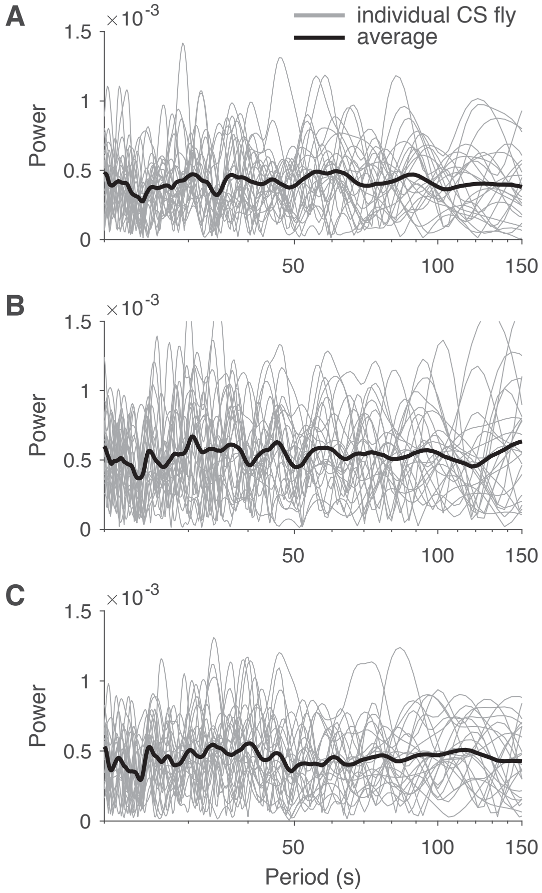


Figure S3. Lomb-Scargle periodograms for *Canton-S* song recordings. (A) From data manually annotated by Kyriacou et al. (3). (B) Automatically segmented data from Stern (3). (C) Automatically segmented data using segmentation parameters from Coen et al. (5). Individual recordings are shown in grey and average over all recordings is shown in black.

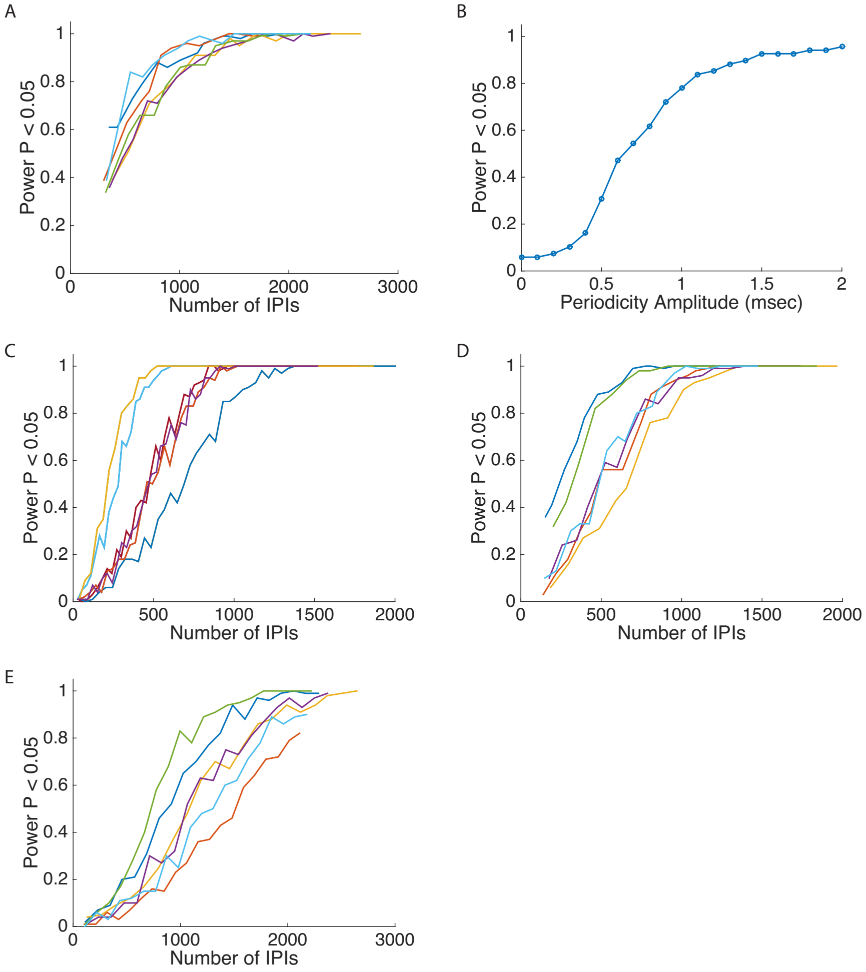


Figure S4. Statistical power analysis under multiple scenarios. (A) Power analysis after ten-second bins of inter-pulse interval data were removed randomly. The plots show the proportion of times out of 100 that periodicity was found between 50-60 sec with P < 0.05 for each of six songs containing more than 10,000 inter-pulse interval events. (B) Dependence of power to detect simulated periodicity on periodicity amplitude. Simulated periodicity of 55 sec with amplitude between 0 and 2 msec was imposed on sixty-eight Canton-S songs containing at least 1000 inter-pulse interval measurements. Power equals the fraction of songs that displayed power between 50 and 60 sec at P < 0.05. (C, D) Simulated periodicity was added to six songs containing at least 10,000 inter-pulse interval (IPI) events in 45 minutes and then only the first 400 seconds of the song were analyzed. One hundred times, inter-pulse interval data were dropped either randomly (C) or 10 sec bins were dropped randomly (D) and Lomb-Scargle periodogram analysis was performed. (E) Power to detect a sawtooth rhythm. Sawtooth periodicity was added to six songs containing at least 10,000 inter-pulse interval (IPI) events in 45 minutes.

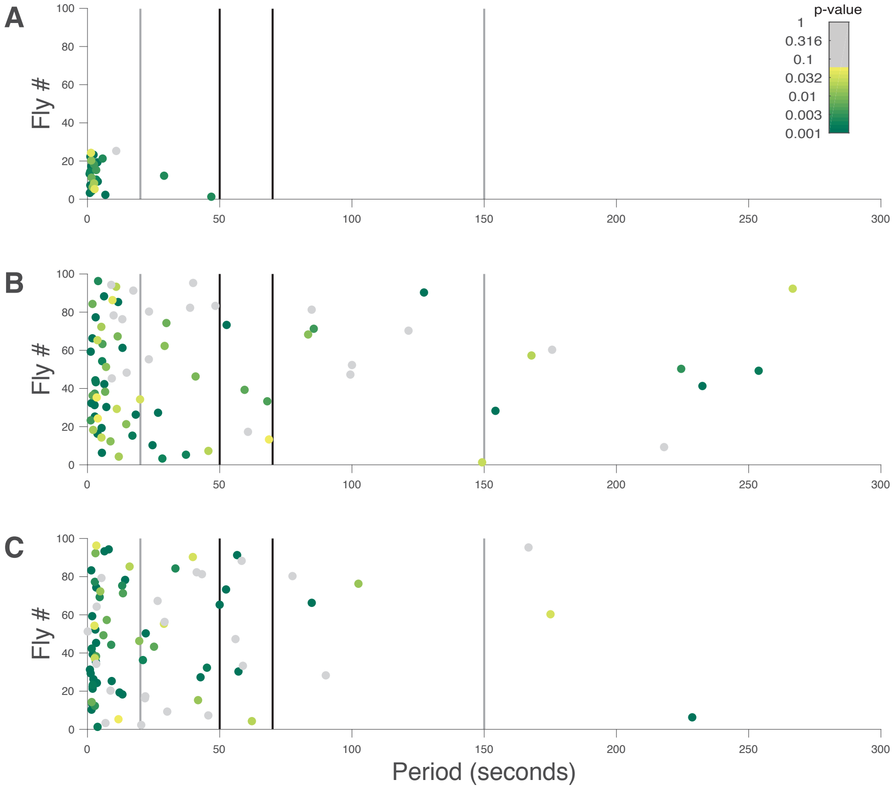


Figure S5. Period of maximum Lomb-Scargle periodogram peaks for inter-pulse interval measurements from multiple individual recordings. (A) Songs manually annotated by Kyriacou et al. (3). (B) Songs automatically segmented in Stern (2). (C) Songs from Stern (2) automatically segmented using parameters defined in Coen et al. (5). In all cases, the vast majority of significant rhythms cluster in the highest frequency range (low period). But both non-significant and significant peaks are distributed widely and apparently at random across the frequency range. In each plot, the 50-70 sec period range is defined by the vertical black lines and the 20-150 sec range defined by Kyriacou et al. (3) is shown with vertical gray lines.

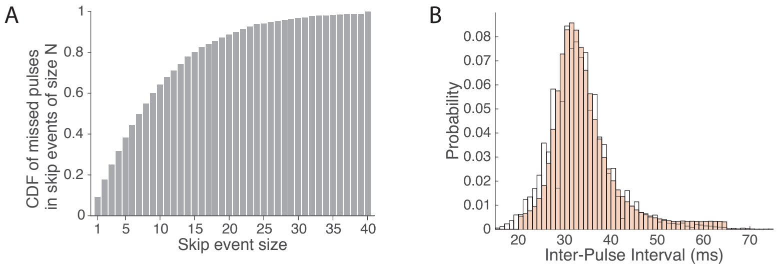


Figure S6. Missed pulse events cause a minor change to the distribution of inter-pulse interval events. (A) Cumulative density function of the number of consecutively missed inter-pulse interval values in the data from Stern (2) illustrates that only 9% of missed pulses were singletons that might alter retained inter-pulse intervals. (B) Histogram of inter-pulse interval data from all *Canton-S* recordings from Stern (2014) in orange and from all manually annotated *Canton-S* recordings from Kyriacou et al. (3) in white. The automatically scored data display a slight excess of inter-pulse interval (IPI) values in the range of approximately 50-65 msec, which are unlikely to significantly alter downstream analysis, as shown by analysis of inter-pulse interval cutoffs in the following panels.

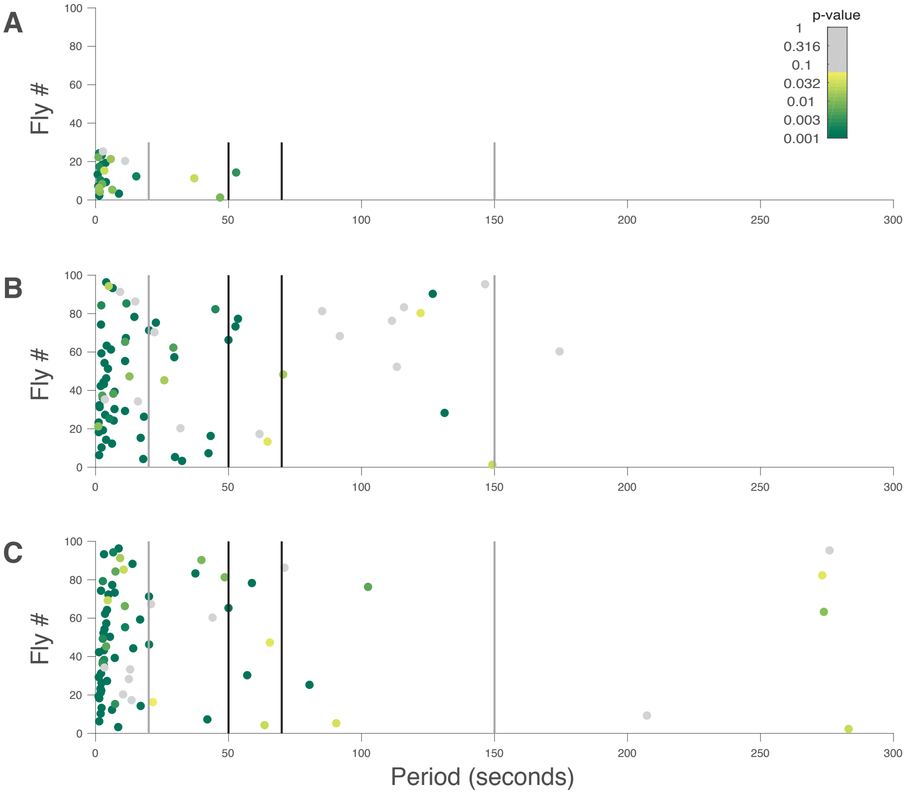


Figure S7. Period of maximum Lomb-Scargle periodogram peaks for inter-pulse interval measurements from multiple individual recordings with an inter-pulse interval cutoff of 55 msec. (A) Songs manually annotated by Kyriacou et al. (3). (B) Songs automatically segmented in Stern (2). (C) Songs from Stern (2) automatically segmented using parameters defined in Coen et al. (5). In all cases, the vast majority of significant rhythms cluster in the highest frequency range (low period). But both non-significant and significant peaks are distributed widely and apparently at random across the frequency range. In each plot, the 50-70 sec period range is defined by the vertical black lines and the 20-150 sec range defined by Kyriacou et al. (3) is shown with vertical gray lines.

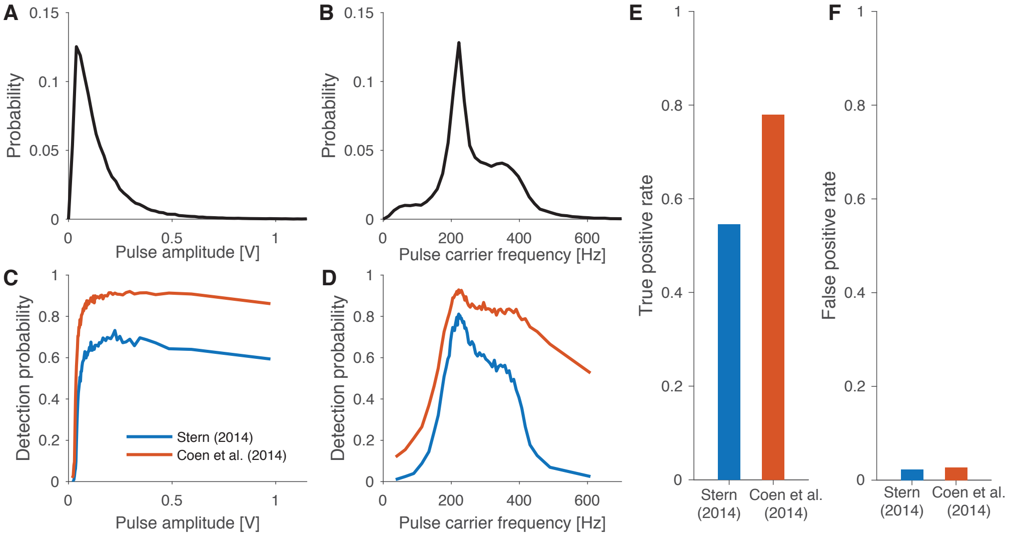


Figure S8. Modification of initialization parameters of FlySongSegmenter influences its performance in detecting pulses. (A, B) Distribution of pulse amplitudes (A) and carrier frequencies (B) for the pulses manually annotated in Kyriacou et al. (3). (C, D) Probability of detecting manually annotated pulses by the automated song segmenter using either the initialization parameters from Stern (2) or Coen et al. (5) versus pulse amplitude (C) or pulse carrier frequency (D). (E, F) True (E) and false (F) positive rate of pulse detection using parameters from Stern (2) and Coen et al. (5) for the pulses manually annotated in Kyriacou et al. (3).

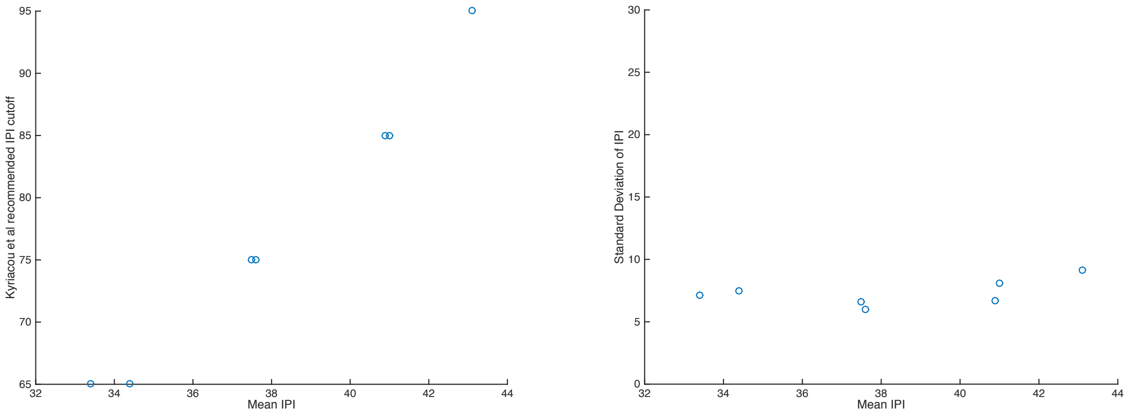
Inter-pulse interval cut-off and temperature control

Under the heading “Problem 2: Inappropriate upper IPI cut-offs and poor temperature control,” Kyriacou et al. (1) state that Stern (2) used an inappropriate upper inter-pulse interval cutoff for some of the songs and that temperature was not controlled during experiments. We address each concern in turn.

Inter-pulse interval cut-off: Kyriacou et al (1, 3) recommended that the IPI cut-off should scale with the mean inter-pulse interval for a genotype. They did not indicate precisely how the cut-off should scale with the mean. In their table S1, they indicated a “more appropriate cutoff” for each genotype without a quantitative description of how this cutoff should be calculated. The mean inter-pulse intervals and standard deviations calculated from all songs with > 1000 IPIs are shown below along with their recommended upper cut-off.

	per^{01}	per^{L}	per^{S}	*D. simulans*	CantonS	CantonS Manual	per^{L} Manual
Mean IPI	41.0	37.6	40.9	43.1	34.4	33.4	37.5
Recommended IPI cut-off	85	75	85	95	65	65	75
Std Dev IPI	8.09	5.99	6.72	9.13	7.46	7.11	6.59

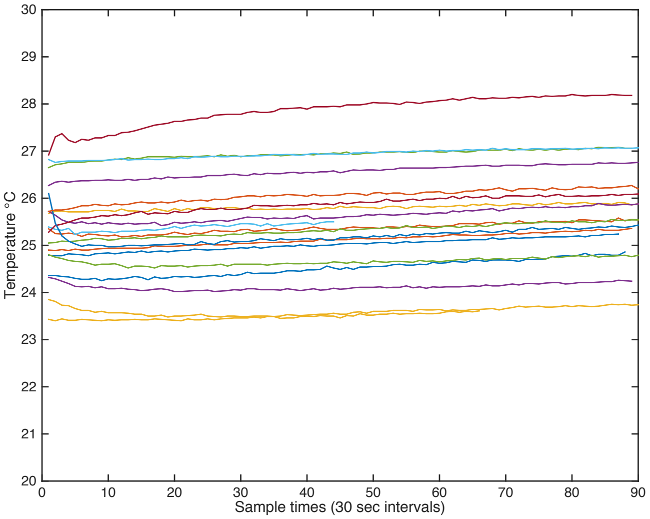
The mean inter-pulse interval varies by less than 10 msec, but the recommended cut-offs vary by 30 msec. The slope of the regression of mean inter-pulse interval and the recommended cut-off is 3.1 ($y = 3.1x – 40$). In essence, Kyriacou et al. assume that the standard deviation in inter-pulse interval increases considerably faster than the mean inter-pulse interval (plot below left). We find, in contrast, that the standard deviation in inter-pulse interval is relatively constant across genotypes ($y = 0.14x + 1.8$ for automated data) (plot below right). Changing the cutoff by the change in the mean, rather than 3X faster than the mean, is justified by these observations.



Even more importantly, however, in the main manuscript we report simulations where we progressively reduced the IPI cutoff for song with simulated rhythms. We find that the upper cut-off can be reduced from 65 ms to at least as low as 25 ms and simulated periodicity can still be detected as long as the song retains at least 1000 inter-pulse interval events. It is unlikely, therefore, that any particular IPI cutoff has any influence on the ability to detect song periodicity.

Temperature: Environmental temperature is known to influence the inter-pulse interval of courtship songs. There is no report that temperature can influence the proposed rhythm in the inter-pulse interval, but Kyriacou et al (1) claimed that the experiments reported in Stern (2) had poor temperature control and that this might cause problems with analysis.

We re-examined the data and found that, indeed, average temperature did vary between recording sessions with a range of approximately 4.3°C. However, within each 45-minute recording session, temperature varied on average with a range of 0.52°C. On average, temperatures in the chambers increased slightly over the course of the recording session, likely due to the heat produced by the electronics. In the plot below, we show the temperature for each experiment shown in a different color over each approximately 45-minute recording.



While these slight differences in temperature over the course of each experiment are expected to have a subtle effect on the inter-pulse interval, it is not clear that song periodicity should *disappear* as a result of these small temperature changes. One might imagine that the periodicity might differ at different temperatures, but the essential point of Stern (2), emphasized by results in this paper, is that periodicity itself could not be detected.

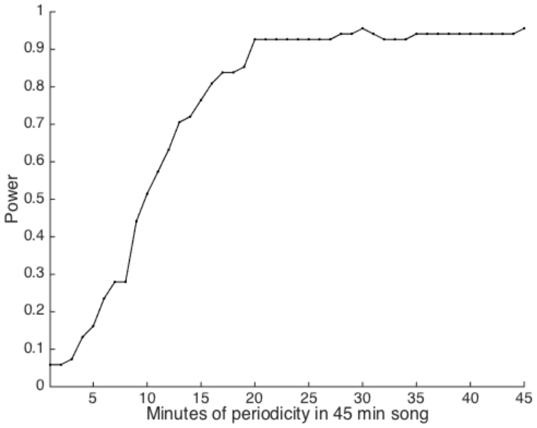
Kyriacou et al also stated that differences in mean IPI should be incorporated into changes in the IPI cutoff. Different experiments were recorded at temperatures that varied by at most ~4°C, although most experiments were recorded at temperatures of between ~25°C and ~26°C. Variation of ~4°C is expected to alter mean IPI by only ~5 msec (4). So, this might justify a change in the IPI cutoff of up to a maximum of 5 msec, which is unlikely to alter any of the statistics substantially. In addition, we showed in the main manuscript that changing the IPI cutoff by up to 40 msec (from 65 msec to 25 msec) has little effect on the ability to detect simulated rhythms, so a small change in the IPI cutoff is unlikely to resolve the question of whether IPI periodicity exists.

Length of courtship

Under the heading “Problem 3: Unrealistic length of courtship,” Kyriacou et al. (1) state that “courtship interactions under natural conditions are brief,” lasting less than 30 sec and therefore question the use of 45 minute recordings of song. (Of course, if courtship really lasted less than 30 sec, then 50-60 sec periodicity could not be detected.) The key reference the authors cite for natural courtships (5) indeed reported that the majority of courtship interactions lasted less than 30 seconds, however, none of the 153 courtship interactions observed in that study ended in copulation. It is possible that most or all of the females studied were not virgin and were unwilling to participate in courtship. Therefore, these data are not relevant to the question of how long courtship between a male and virgin female persists in nature.

Kyriacou et al. (1) further question the use of 45 minute recordings because circadian rhythms can dampen quickly, citing (6). Reference (6) reports on dampening of circadian rhythms during real-time luminescence recording from cultured explanted rat superchiasmatic loci over the course of approximately 10 days. One can imagine multiple reasons why cultured cells would display a dampened rhythm over 10 days. It is not clear how this is relevant to a presumptive song rhythm over a roughly 45-minute time span.

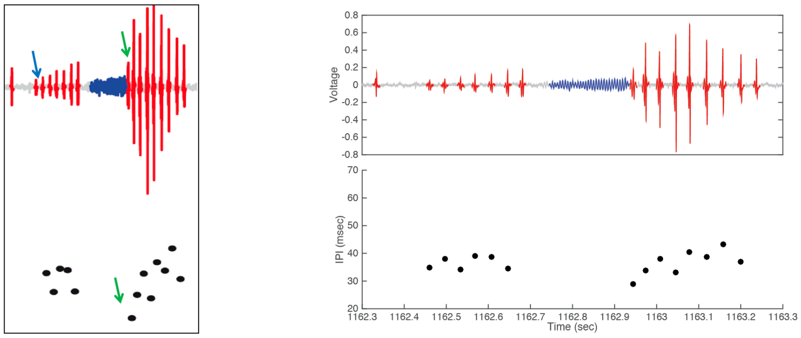
Nonetheless, we decided to investigate this issue more closely. First, we examined the power to detect periodicity in songs if the periodicity was present for only the first N minutes of the song. Periodicity was imposed on the first N minutes of 45 minute recordings for 68 Canton-S songs that contained more than 1000 inter-pulse interval measurements and the average probability of detecting this periodicity with LS periodogram analysis is reported as power in the plot below. We retained power greater than 0.8 as long as periodicity persisted for at least the first 16 minutes. In addition, the probability of detecting periodicity rose above random (P = 0.05) with as little as three minutes of periodicity. Thus, it is extremely unlikely that we would have failed to have detected periodicity in the song recordings as long as periodicity persisted for more than a few minutes.



Furthermore, if we perform power analysis only on songs 400 sec long, then we retained power of > 0.8 as long as these short songs contained at least 1000 inter-pulse interval events (Fig. S4a), even when pulses were dropped in 10-sec bins (Fig. S4b). Thus, there is no evidence that the length of courtship recordings generated data that are biased against detecting courtship rhythms.

Reanalysis of Stern's primary matlab song records

Kyriacou et al. (1) observed an apparent error (blue arrow below) in the calling of an inter-pulse interval in Figure 1b of Stern (2) and report this in Fig S2 of their paper. Figure 1b in Stern (2), reproduced below on left, was derived from experiment PS_20130625111709_ch3, sample points approximately 1162.3 sec to 1163.3 sec. We have re-examined the original data and find that the apparently missing inter-pulse interval is in fact found in the csv file that was provided with the original manuscript, but was inadvertently deleted during construction of the figure. We have replotted the data below on the right.



References

1. Kyriacou CP, Green EW, Piffer A, Dowse HB, Takahashi JS (2017) Failure to reproduce period-dependent song cycles in Drosophila is due to poor automated pulse-detection and low-intensity courtship. *PNAS*. doi:10.1073/pnas.1615198114.
2. Stern DL (2014) Reported Drosophila courtship song rhythms are artifacts of data analysis. *BMC Biol*.
3. Kyriacou CP, van den Berg MJ, Hall JC (1990) Drosophila courtship song cycles in normal and period mutant males revisited. *Behav Genet* 20(5):617–644.
4. Peixoto AA, Hall JC (1998) Analysis of temperature-sensitive mutants reveals new genes involved in the courtship song of Drosophila. *Genetics* 148(2):827–38.
5. Gromko M, Markow T (1993) Courtship and remating in field populations of Drosophila. *Anim Behav* 45:253–262.
6. Yamazaki S, Takahasi JS, Takahashi JS (2005) Real-Time Luminescence Reporting of Circadian Gene Expression in Mammals. *Methods Enzymol* 393:288–301.