Reported *Drosophila* courtship song rhythms remain data analysis artifacts.

Authors:

David L. Stern[1], Jan Clemens[2], Philip Coen[4], Adam J. Calhoun[2], John B. Hogenesch[5], Ben Arthur[1], and Mala Murthy[2,3]

Affiliations:

[1] Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA

[2] Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA

[3] Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

[4] University College London, Gower St, Bloomsbury, London WC1E 6BT, UK

[5] Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH 45229

Correspondence: sternd@janelia.hhmi.org

Abstract

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 performed automated segmentation of songs, but failed to detect the proposed periodicity (1, 2). Kyriacou et al. (3) report that we failed to detect song rhythms because i) our flies did not sing enough and ii) our segmenter did not identify song pulses accurately. They 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, as well as the automatically segmented data, provide no evidence for either the existence of song rhythms or song periodicity differences between genotypes. Furthermore, we have re-examined our methods and analysis and find that our methods did not prevent detection of putative song periodicity. We therefore conclude that previous positive reports of song rhythms most likely resulted from inappropriate statistical analyses.

Significance statement

Previous studies have reported that male vinegar flies sing courtship songs with a periodic rhythm of approximately 60 seconds. 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 1) our flies did not sing enough and 2) our software for detecting song was flawed. They claimed that they could detect rhythms in song annotated by hand. We report here that we cannot replicate their observation of rhythms in the hand-annotated data and that our original methods were not biased against detecting rhythms. We conclude that the original findings likely resulted from errors in the statistical analysis of songs.

Introduction

When a male vinegar fly (*Drosophila melanogaster*) encounters a sexually receptive female vinegar fly, he initiates a complex series of behaviors including the production of elaborate courtship songs. Males produce songs, containing pulses and hums (or sines), via unilateral wing vibration (Fig. 1a). Quantitative assessment of these songs reveals that every parameter—including the amplitude and frequency of pulses and sines and the timing of individual pulse and sine events—displays extensive variation within a single bout of singing (1, 2, 4–8). Like humans during conversation, *Drosophila* males modulate the content and amplitude of their song based on sensory feedback from their communication partner (4, 5).

The time between pulse events in a single train, the inter-pulse interval, varies extensively within an individual male's song (Fig. 1b)(1, 9). Visual inspection of these songs reveals that the mean inter-pulse interval varies over time (Fig. 1b). This observation was first made in 1980 by Kyriacou and Hall (10), where they claimed that this variation cycled with a periodicity of about 55 sec and was controlled, in part, by the *period* gene, a gene with a key role in circadian rhythms (11). Later papers presented evidence that evolution of a small sequence within the *period* protein between *D. melanogaster* and *D. simulans* was responsible for species-specific differences in cycling of the inter-pulse interval (11–14). This series of reports attracted considerable interest because it implicated the *period* gene in ultradian rhythms, in addition to its well-known role in circadian rhythms(15), and because it represented an impressive example of how genetic evolution caused behavioral evolution.

Despite this progress, the molecular mechanisms underlying these phenomena remain unknown. In an effort to further advance study of these rhythms, previously we analyzed many

songs from both *D. melanogaster* and *D. simulans* and searched for the rhythms using sensitive methods to detect periodicity in time series data (1). We failed to find evidence for the song rhythms. We were mindful, however, that Kyriacou and Hall had argued that the presence or detectability of the rhythms was sensitive to assay conditions and methods of analysis (16). One of us, therefore, decided to attempt to replicate the methods of Kyriacou and Hall as closely as possible, but, again, song rhythms could not be detected (2).

To understand this failure to replicate earlier studies, it is critical to consider the precise methods of analysis used by previous papers. To enable quantitative analysis of the inter-pulse interval variation by time series analysis, which usually requires equally spaced samples, Kyriacou and Hall (10) had binned data into 10 sec intervals and interpolated values for missing bins. The choice of 10 sec bins was never justified and, in a previous paper, we reported that binning the data, together with the analysis of relatively short songs, artificially creates peaks in spectrogram analysis that fall within a relatively narrow frequency range, corresponding approximately to the frequency range originally reported for the periodicity (2). This would, at first, appear to support the claims of earlier papers. However, fewer than 5% of these peaks reached a nominal significance level of $p<0.05$ (four of 149 songs, Fig. 3a of (2)), strongly suggesting that these peaks represent signals that cannot be distinguished from noise. Moreover, the clustering of these non-significant peaks in a specific frequency range is an artifact of how the data were analyzed, namely binning data from short songs. Nonetheless, these are the peaks that previous papers had utilized for further statistical analyses of different genotypes.

Kyriacou et al. (3) (subsequently "the authors") have recently published a paper that questions our previous conclusions. Here we focus on three major assertions that they claim call our conclusions into doubt. First, we examine the author's central claim that manual analysis, in contrast to automated analysis, of songs reveals genotype-specific song rhythms. In agreement with our original findings, we find no compelling evidence that song rhythms exist and re-analysis of their manually-annotated data provides no statistical support for genotype-specific rhythms. We also find no evidence for genotype-specific rhythms in a new larger dataset. Second, we examined the authors' claim that the original recordings contained insufficient data to detect rhythms and find that this claim is not supported by simulation studies. Third, these observations imply that the authors' concerns about the observed false negative and false positive rate of the automated song segmenter are not relevant. We identify the major biological sources of false negative events and illustrate that minor modifications to initialization parameters can improve segmenter performance. The authors also raised a number of minor concerns—such as how to choose an appropriate inter-pulse 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 core questions raised and therefore we have addressed these concerns (which are also unsupported by correct data analysis methods) in a supplemental file (Supplementary Data).

Results

No support for the claim that manual song segmentation reveals genotype-specific song rhythms.

The authors' core finding is that different genotypes displayed different periodic rhythms of the inter-pulse interval. This is also the most important discovery reported in earlier papers on this subject (11–13, 17). In their manual analysis of our data, the authors focused on a comparison of recordings from a wild-type strain, *Canton-S*, and a strain carrying a specific mutation in the *period* gene, per^L. Flies homozygous for per^L display circadian rhythms that are longer than normal (15) and earlier papers have reported that per^L confers longer periods on the inter-pulse interval rhythm (10–13). The authors reported that when they manually identified pulse events in a subset of some of our recordings, they detected a difference in the mean song period between *Canton*-S and per^L, but they did not detect this difference when using the data generated by automated song segmentation.

It is important to clarify precisely what the authors measured and compared in this test. They have claimed that the inter-pulse interval varies, on average, with a regular periodicity. Therefore, it should be possible to detect this rhythmicity with appropriate methods of periodogram analysis. We have previously employed Lomb-Scargle periodogram analysis (18–20) because this method does not require evenly spaced samples and the authors also adopted this method. For example, the Lomb-Scargle periodogram analysis of the time series shown in Fig. 1b is shown in Fig. 1c. In this case, despite the obvious variation in inter-pulse interval values observed in Fig. 1b, there is no significant periodicity observed between 20 and 150 sec.

The authors specified that wild-type *D. melanogaster* song rhythms are expected to occur with a periodicity between 20 and 150 sec. This is a much wider range of periodicities than the approximately 50 – 60 sec periods originally reported by Kyriacou and Hall (10). The authors do not clarify why they chose this frequency range for analysis, but increasing the width of the periodicity window from 50-60 sec to 20-150 sec increases the probability of detecting spurious significant periods. However, even given this wide frequency range, we observed that only four of the 25 *Canton-S* songs manually annotated by the authors and three of 25 automatically segmented songs contained power that reached a significance level of $P < 0.05$. (When we follow Kyriacou & Hall's (10) protocol of binning the data in 10 sec bins, these values decline to 0 of 25 manually annotated and 1 of 25 automatically segmented songs.) Because so few songs produce significant periodicity in the focal frequency range, the authors therefore followed the same protocol that they have advocated in earlier papers, which is to identify the peak in the periodogram that has maximum power in this frequency range (regardless of significance or of the power of signals outside this range) and use this as the best estimate of the song rhythm for each fly.

This is an unorthodox approach to data analysis. The typical interpretation of non-significant regions of a periodogram is that there is no signal in this frequency range that can be distinguished from noise. One interpretation of the authors' approach is that they have implicitly assumed that current methods of periodic signal detection are under-powered to detect song rhythms. This seems unlikely, since we have found that simulated song rhythms can be detected with high confidence ((1, 2) and see below) even for very noisy rhythms (1). Nonetheless, if we assume that the song rhythms have unusual properties and that

periodogram analysis is underpowered to detect these rhythms, then we should observe that songs tend to display nearly-significant periodicity. In fact, we observe that 72% of p-values are greater than p = 0.2 (Fig. S1). There is therefore no evidence for an excess of periodicity with nearly significant p-values.

An alternative possible interpretation of the authors' inclusion of non-significant periodogram peaks is that the signal to noise of the periodicity is extremely low. An analogue 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 the results for each genotype. These plots are essentially flat, suggesting that there is no signal hidden in the fluctuations of individual periodograms (Fig. S2).

Given these observations, further analysis of these data seems unwarranted. However, The authors proceeded to compare the *Canton-S* and per^L genotypes and found that the manually-annotated data showed a statistically-significant difference in the mean period, although the automatically segmented data did not (the authors' Figure 3d). This is the key result of their paper, which appears to both corroborate findings reported in earlier papers and justify manual segmentation of songs. We downloaded the manually annotated data provided by the authors and for each song identified the peak in the periodogram of maximum power falling between a period of 20 and 150 sec. In contrast to the result reported by the authors, we found that the average period with maximum power (most of which were not significant) was not significantly different between the genotypes *Canton-S* and per^L (Figure 1d). We have no explanation for this discrepancy between our statistical analysis and theirs.

Since the authors did not provide a biological or quantitative justification for the particular frequency ranges examined in any study, we wondered whether the results were sensitive to the precise frequency range examined. We found that the test statistic was extremely sensitive to the precise frequency range selected (Fig 1e). The vast majority of frequency windows do not generate a statistically significant difference between the genotypes (Fig. 1e,g), and false discovery rate correction for multiple testing (21, 22) yields no frequency ranges with significant results (Fig 1f,g).

This analysis reveals that there is no support for the specific results reported by the authors. Furthermore, there is no statistical support for defining song cycle periods as occurring within any particular window. Most importantly, our analysis indicates that secondary genotype-specific analysis of non-significant 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 using these methods, 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.

New data provide no support for the claim that different genotypes confer different song periodicities

The previous analysis strongly suggests that the statistically significant results reported by the authors are artifacts of improper data analysis. However, we thought it may be worth taking their observation at face value as a preliminary result and testing directly whether we

can detect genotype specific song rhythms in a new, expanded data set. We therefore recorded new song from 33 *Canton-S* males and 34 *period*^{L} males. Following the authors' procedure of analyzing the strongest periodogram peak in the frequency range of 20-150 s, we found no significant difference between these genotypes (Fig. 1h). We then compared test statistics across a wide set of frequency ranges, as described in the previous section. We identified some frequency ranges that yielded significant results in the predicted direction (Fig. 1i), 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 (Figure 1j). Second, multiple frequency ranges support the *opposite* conclusion, that *Canton-S* rhythms are slower than *period*^{L} rhythms (Figure 1k). Third, these frequency ranges only partially overlap with the ranges found for the original dataset (c.f. Figures 1e & 1i). 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.

No support for the claim that low-intensity courtship song contained insufficient data to detect song rhythms

While 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 strong claims made by the authors because we also find no support for these claims. The authors claimed that rhythms can be detected only in songs produced by "vigorously" singing males. They write that "sporadic songs could not possibly provide any test for song cycles." However, they provided

neither a biological nor quantitative justification for requiring that singing must be “vigorous” to detect song rhythms. It is not clear if they are claiming that rhythms can be detected only in songs with many pulses or that only flies that sing songs with many pulses produce rhythms. We therefore evaluate each alternative interpretation in turn.

We previously investigated songs from 45 minute courtship recordings that contained at least 1000 inter-pulse interval measurements (2). The authors claimed that for songs 45 minutes long, detection of rhythms requires song with more than 5000 inter-pulse interval measurements. To examine this claim, we performed a statistical power analysis using songs with variable numbers of inter-pulse 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. We started with five 45-minute recordings of *Canton-S* from reference (2) that contained more than 10,000 inter-pulse interval measurements. None of these five songs yielded statistically significant power in the frequency range between 50 and 60 Hz (the range originally defined to contain rhythms (10)) and one song produced a marginally significant peak at 31.7 Hz ($P = 0.04$), which falls between 20 and 150 Hz (the range recently employed by the authors to search for peaks in the power spectra). Figure 2b and 2c illustrate the inter-pulse interval data and periodogram for one of these songs. Therefore, these songs do not contain strong periodicity in the range predicted by the authors and can serve as a useful 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 inter-pulse interval reported rhythms with a mean period of 55 sec and an amplitude of approximately 2 ms (10). Therefore, we imposed a

55 second rhythm with an amplitude of 2ms on the five songs containing > 10,000 inter-pulse interval measurements (Fig. 2a). In a previous power analysis, we found that Lomb-Scargle periodogram analysis detected simulated periodicity in songs even when the periodicity was extremely noisy; a simulated rhythm could be detected greater than 80% of the time when the signal to noise ratio was above one (1). Since the authors claimed that most of our songs did not contain sufficient data to detect rhythms, here we extend the power analysis by removing data points to determine whether the songs analyzed previously contained sufficient data to detect putative rhythms.

We detected the simulated 55 sec rhythm in all five songs with P-values << 0.001 (example shown in Fig. 2d, e). (These data were not pooled into 10 sec bins.) 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 sec 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. 2f,g; summary statistics shown in Fig 2h,i). That is, as long as songs contained at least 1000 inter-pulse 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 sec of songs (Fig. S3). Furthermore, periodicity could be detected with power greater than 0.8 when the amplitude of simulated periodicity was greater than at least 1 msec (Fig. 2j). The authors' claim that only songs with > 5000 inter-pulse interval events

can be used to detect periodic cycles is not supported by this simulation analysis. Instead, our initial choice of 1000 inter-pulse intervals (2) appears to be a reasonable threshold to detect putative rhythms with high sensitivity. It is worth re-emphasizing that all of these results were generated without binning the data, although previous papers (3, 10, 16) have repeatedly advocated averaging inter-pulse interval data in 10 sec bins. This is inadvisable because, as we showed earlier (2), binning only reduces the significance of periodogram peaks.

It is harder to evaluate the interpretation that only males that sing robustly *produce* rhythms. As we demonstrated above, we find no compelling evidence for inter-pulse interval rhythms in our recordings that contain far more pulses than the authors defined as the minimum for "vigorous courtships". This suggests that even the most vigorously singing males do not produce rhythms. However, previous studies have historically included non-significant periodogram peaks in their down-stream analysis (3, 16), so they may claim that our findings are not relevant given their methods of analysis. We evaluated this procedure earlier and find no support for this analysis method.

Previous papers (3, 16) have also claimed that song rhythms can be detected only in the first few minutes of courtship. One interpretation of this claim is that only "robustly" singing males produce rhythms, but they do so only in the first few minutes of courtship. This claim was evaluated in Stern (2) and he found no compelling evidence for song rhythms in the first five minutes of recordings. Furthermore, the authors examined our songs that met their thresholds for inclusion and found no evidence for rhythms in the first 400 sec of our segmented song. We have repeated this analysis and agree that there is no evidence for periodic cycles in the first

400 sec of our recordings, which is consistent with the lack of evidence for rhythms throughout the rest of each song.

We conclude that, contrary to the authors' strong claim that songs must contain more than 5000 inter-pulse interval events to allow detection of song cycles, songs containing at least 1000 inter-pulse 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, as long as approximately 1000 inter-pulse intervals remain. We found previously (2) that periodicity similar to the putative song cycles cannot be identified in the vast majority of automatically segmented songs (e.g. only two of 68 Canton-S recordings reported in Fig. 4 of Stern (2) exhibit p-values < 0.05 in the relevant periodicity range) and we showed above that there is no compelling evidence for periodicity in the manually annotated song. One critical point is that 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 (Fig. S4; Fig. 4 of Stern (2)). In addition, no songs are significant after correcting for multiple comparisons (Fig. 1). All together, these results imply that the few *nominally statistically* significant results that can be found do not carry *biological* significance and instead reflect random fluctuations in the inter-pulse interval.

No support for the claim that the automated fly song segmenter biased the results

A core claim of the authors is that the fly song segmenter displayed a low true positive rate (the segmenter fails to detect some actual song pulses) and an elevated false positive rate (the segmenter classifies some noise as song pulses). They suggest that these incorrect pulse event assignments could bias estimation of the mean inter-pulse interval and therefore decrease the signal-to-noise of the periodic cycle, making it difficult to detect a periodic signal. In principle, a large sample of incorrect calls could bias results, so we investigated whether this was the case for our prior analyses. We used the authors' manually-annotated dataset first to investigate the potential for bias and second to evaluate performance of the automated segmenter.

When a single pulse event is not detected, the inter-pulse interval is then calculated as the sum of the two neighboring real intervals. On average, this is approximately double the average inter-pulse interval. The average inter-pulse interval for the *Canton-S* recordings reported in Stern (2) is approximately 35 msec with a standard deviation of approximately 7 msec. Therefore, skipping a single pulse event is expected to result in inter-pulse interval measurements of approximately 70 msec, but with considerable variance. Following Kyriacou and Hall (16), Stern (2) employed a heuristic threshold of 65 msec to reduce the number of spurious inter-pulse interval values. Therefore, in the specific case when a single pulse in a train is missed, approximately one third of the incorrectly scored doublet inter-pulse interval measurements would be shorter than 65 msec 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 inter-pulse interval

measurements that are excluded by the 65 ms threshold. Using the authors' manually annotated data, we found that only 9% of the pulses missed by automated segmentation were singletons (Figure 3a). These incorrect inter-pulse intervals contribute to a slight excess of inter-pulse intervals with high values (Figure 3b). Lowering the inter-pulse interval threshold would, therefore, remove most or all spurious inter-pulse intervals. Since our power analysis, discussed above, revealed that periodogram analysis was robust to random removal of inter-pulse interval events, as long as songs still contained at least 1000 values, loss of a small number of inter-pulse intervals is not expected to hamper detection of rhythms. After reducing the inter-pulse interval threshold to 55 msec, we still found no compelling evidence for significant periodicity in the original data (Fig. S5). Therefore, we explored the effect of reducing the inter-pulse interval cutoff on the statistical power to detect rhythms in songs containing simulated periodicity. In this case, we used all 68 *Canton-S* songs from Stern (2) and retained for analysis only those songs that contained at least 1000 inter-pulse interval measurements after imposing the new inter-pulse interval threshold. We explored a range of cutoff values from 25 to 65 msec. We found that we could detect the simulated rhythm in most songs with at least 1000 inter-pulse interval measurements remaining after thresholding, even when the threshold was as low as 25 msec (Fig. 3e-g). Therefore, there is no evidence that pulses missed by the automated song segmenter or the specific inter-pulse interval threshold used in Stern (2) prevented detection of song rhythms.

Although detection of putative song rhythms is robust to dropped pulses in songs with at least approximately 1000 inter-pulse 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 prior to 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 & Hall (10). This larger chamber with one microphone had reduced sensitivity compared to 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 approximately 35% of the missed pulses (Fig. 4a, 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 non-song noises, like grooming, and a user can set parameters in the segmenter to attempt to exclude these non-song 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 approximately 42% of the missed pulses (Figure 4b,d). Using the same software with different parameters (from Coen et al. (5)) recovers many of these high-frequency pulses without substantially increasing the false positive rate (Figure 4c-f).

Since the recordings used in Stern (2) provided sufficient data to detect putative rhythms in simulations, he did not invest additional effort to optimize parameters to increase

the true positive rate of pulse detection. 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 data set that was biased against detection of song rhythms. 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 inter-pulse interval in *Drosophila* courtship song even in the songs manually annotated by the authors and used as evidence for periodicity in their paper. While 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). Previously, we offered one explanation for how previous authors may have convinced themselves that they had detected song rhythms. 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 almost none of these peaks reached statistical significance, previous authors have consistently accepted these peaks as “signal” and performed statistical analyses to compare the peaks between genotypes. This is one possible explanation for the results published in earlier papers. However, there may be a more prosaic explanation for both the initial discovery and the repeated reporting of periodic song cycles.

Every fly produces highly-variable inter-pulse intervals. In addition, a running average of these data reveals that the average inter-pulse interval cycles in a pattern (Fig. 1b) similar to

the temporally-binned data first reported by Kyriacou and Hall (10). There is no debate about this observation. The claim in dispute is that the average inter-pulse 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 inter-pulse interval cycled regularly. Indeed, all reports that have claimed to identify the rhythm have reported that some or most flies do not sing rhythmically (2). The flies that did not exhibit the rhythm were discarded from statistical analysis: approximately 10% discarded in 1980 (10), approximately 25% discarded in 1984 (11), and approximately 30% discarded in 1991 (12). The discoverers have repeatedly claimed that the phenomenon is extremely sensitive to recording conditions and the way data are analyzed and they have required the application of an ever-increasing list of conditions to detect the rhythm (3, 16, 17). We find, however, that we cannot detect the rhythm even with application of the most rigorous set of conditions and that an exploration of parameter space around these specified conditions also fails to yield evidence for rhythms.

A much simpler explanation for the extraordinary within-fly variation in the inter-pulse interval and in the mean inter-pulse interval is that male flies respond to ever-changing cues during courtship and modulate their inter-pulse interval to optimize their chances of mating. There is now considerable evidence that individual *Drosophila* males modulate specific aspects of their courtship song in response to feedback from females, including the transition between sine and pulse song (5) and the amplitude of pulse song (4). There is additional evidence for modulation of the carrier frequency of sine song (1). We hypothesize that male flies also modulate their inter-pulse interval in response to specific external cues.

Earlier papers have claimed repeatedly that the best evidence for the existence of periodicity in the inter-pulse interval is that flies with different genotypes and flies of different species, especially flies carrying mutations in the *period* gene, display different inter-pulse interval cycles (3, 10–13, 16). We cannot replicate this result from the authors' paper using the same data. In addition, it is important to note that all "statistically significant" results from earlier papers are derived mainly from non-significant peaks in periodogram analysis and from relatively small sample sizes in (usually fewer than 10 flies of each genotype), so it is questionable whether these derivative statistics are valid. Nonetheless, we decided to take the authors' latest apparent replication of the difference between per^{L} and *Canton-S* as a preliminary result—which is one appropriate way to treat such non-significant signals—and repeated the experiment. We found no difference between the genotypes with this new sample. We are forced to conclude that previous apparent replications 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. At this time, a conservative assessment of the problem is that *Drosophila* courtship song rhythms cannot be detected and previous reports of such rhythms, and of genotype-specific effects on these rhythms, resulted from statistical analysis artifacts.

Conclusions

While this controversy has revolved, for decades (2, 23–27), around statistical analysis of one particular type of data, the authors have raised an alarm about a much broader issue: whether or not automated methods of behavioral segmentation should be trusted at all. We

believe their alarm is both unfounded and potentially detrimental to progress in the fields of behavior, ethology, neuroscience, and evolution, at the least. Automated segmentation of behavior is a challenging problem, but recent years have seen the introduction of many novel and powerful methods of behavior segmentation (28–35).

There are two major reasons that many scientists are adopting automated segmentation. First, automated segmentation allows analysis of far larger datasets than does human-based annotation. Second, automated segmentation provides unbiased and repeatable assessment of behavior. Both of these advantages of automated segmentation have allowed the discovery and detailed study of many subtle aspects of behavior. Of course, it is important to perform ground truthing of automated segmenters, but the authors seem concerned that segmenters report *any* false positive and false negative events.

Any segmentation process, including manual segmentation, carries both a false positive and false negative rate. The optimal way to overcome this reality is to collect a lot of data with a low false positive rate. A large body of data obviously overcomes even a considerable false negative rate and the central tendency of a large collection of true events will swamp a few false positives. In addition, as we have illustrated, it is often possible through simulation to determine whether the existing false positive and negative assignments influence a test's sensitivity to detect a phenomenon. We have shown that, in this case, they do not. The automated song segmenter could have performed almost ten times worse and not influenced our ability to detect periodic signals, should they have existed. There are many phenomena in nature where it is not possible to capture every event. Faith in the reality of a phenomenon comes from many independent observations, even if observations are sampled sparsely. That

is, we would rather have sparse data from many flies, than perfect data from a small number of flies. In this case, we collected a lot of data from many flies and still we could not detect song rhythms. The path of progress in studies of behavior therefore lies in the direction of increased automated segmentation, not in a return to the limited and potentially biased methods of manual segmentation.

Acknowledgements

We thank Elizabeth Kim for recording the new samples of flies.

Software and data availability

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 is available at https://www.janelia.org/lab/stern-lab/tools-reagents-data.

References

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.

479 2. Stern DL (2014) Reported Drosophila courtship song rhythms are artifacts of data
480 analysis. *BMC Biol*.
481 3. Kyriacou CP, Green EW, Piffer A, Dowse HB, Takahashi JS (2017) Failure to reproduce
482 period-dependent song cycles in Drosophila is due to poor automated pulse-detection
483 and low-intensity courtship. *PNAS*. doi:10.1073/pnas.1615198114.
484 4. Coen P, Xie M, Clemens J, Murthy M (2016) Sensorimotor Transformations Underlying
485 Variability in Song Intensity during Drosophila Courtship. *Neuron* 89(3):629–644.
486 5. Coen P, et al. (2014) Dynamic sensory cues shape song structure in Drosophila. *Nature*.
487 doi:10.1038/nature13131.
488 6. Ding Y, Berrocal A, Morita T, Longden KD, Stern DL (2016) Natural courtship song
489 variation caused by an intronic retroelement in an ion channel gene. *Nature*
490 536(7616):329–332.
491 7. Shirangi TR, Wong AM, Truman JW, Stern DL (2016) Doublesex Regulates the
492 Connectivity of a Neural Circuit Controlling Drosophila Male Courtship Song. *Dev Cell*
493 37(6):533–544.
494 8. Shirangi TR, Stern DL, Truman JW (2013) Motor Control of Drosophila Courtship Song.
495 *Cell Rep* 5(3):678–686.
496 9. Bennet-Clark HCC, Ewing AW, Bennet-Clark HCC (1968) The courtship songs of
497 Drosophila. *Behaviour* 31(3):288–301.
498 10. Kyriacou CP, Hall JC (1980) Circadian rhythm mutations in Drosophila melanogaster affect
499 short-term fluctuations in the male's courtship song. *PNAS* 77(11):6729–6733.
500 11. Zehring WA, et al. (1984) P-element transformation with period locus DNA restores

rhythmicity to mutant, arrhythmic Drosophila melanogaster. *Cell* 39(2 Pt 1):369–376.

12. Wheeler DA, et al. (1991) Molecular transfer of a species-specific behavior from Drosophila simulans to Drosophila melanogaster. *Science* 251(4997):1082–5.

13. Kyriacou CP, Hall JC (1986) Interspecific genetic control of courtship song production and reception in Drosophila. *Science (80-)* 232:494–497.

14. 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.

15. Konopka RJ, Benzer S (1971) Clock Mutants of Drosophila melanogaster. *Pnas* 68(9):2112–2116.

16. Kyriacou CP, Hall JC (1989) Spectral analysis of Drosophila courtship song rhythms. *Anim Behav* 37:850–859.

17. 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.

18. Lomb NR (1976) Least-squares frequency analysis of unequally spaced data. *Astrophys Space Sci* 39(1964):447–462.

19. Scargle JD (1982) Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys Journal, Part 1* 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(2):178–201.

21. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 57(1):289–300.

22. Colquhoun D, London C (2014) An investigation of the false discovery rate and the misinterpretation of P values. *R Soc Open Sci* 1:1–15.

23. Crossley SA (1989) On Kyriacou & Hall's defence of courtship song rhythms in Drosophila. *Anim Behav* (Crossley 1986):861–863.

24. Kyriacou CP, Hall JC (1988) Comment on Crossley's and Ewing's failure to detect cycles in Drosophila mating songs. *Anim Behav* (February):6733.

25. Ewing AW (1988) Cycles in the Courtship Song of Male Drosophila-Melanogaster Have Not Been Detected. *Anim Behav* 36:1091–1097.

26. Logan IG, Rosenberg J (1989) A referee's comment on the identification of cycles in the courtship song of Drosophila melanogaster. *Anim Behav* 37:860.

27. Crossley SA (1988) Failure to Confirm Rhythms in Drosophila Courtship Song. *Anim Behav* 36(Shorey 1962):1098–1109.

28. Berman GJ, Choi DM, Bialek W, Shaevitz JW (2013) Mapping the structure of drosophilid behavior. *Arxiv*:1–22.

29. Kabra M, Robie AA, Rivera-Alba M, Branson S, Branson K (2013) JAABA: interactive machine learning for automatic annotation of animal behavior. *Nat Methods* 10(1):64–67.

30. Branson K, Robie AA, Bender J, Perona P, Dickinson MH (2009) High-throughput ethomics in large groups of Drosophila. *Nat Methods* 6(6):451–457.

31. Dankert H, Wang L, Hoopfer ED, Anderson DJ, Perona P (2009) Automated monitoring and analysis of social behavior in Drosophila. *Nat Methods* 6(4):297–303.

32. Wiltschko AB, et al. (2015) Mapping Sub-Second Structure in Mouse Behavior. *Neuron*

545 88(6):1121–1135.

546 33. Gomez-Marin A, Partoune N, Stephens GJ, Louis M (2012) Automated tracking of animal
547 posture and movement during exploration and sensory orientation behaviors. *PLoS One*
548 7(8). doi:10.1371/journal.pone.0041642.

549 34. Stephens GJ, Johnson-Kerner B, Bialek W, Ryu WS (2008) Dimensionality and dynamics in
550 the behavior of C. elegans. *PLoS Comput Biol* 4(4). doi:10.1371/journal.pcbi.1000028.

551 35. Broekmans OD, Rodgers JB, Ryu WS, Stephens GJ (2016) Resolving coiled shapes reveals
552 new reorientation behaviors in C. elegans. *Elife* 5(September):1–17.

553

554

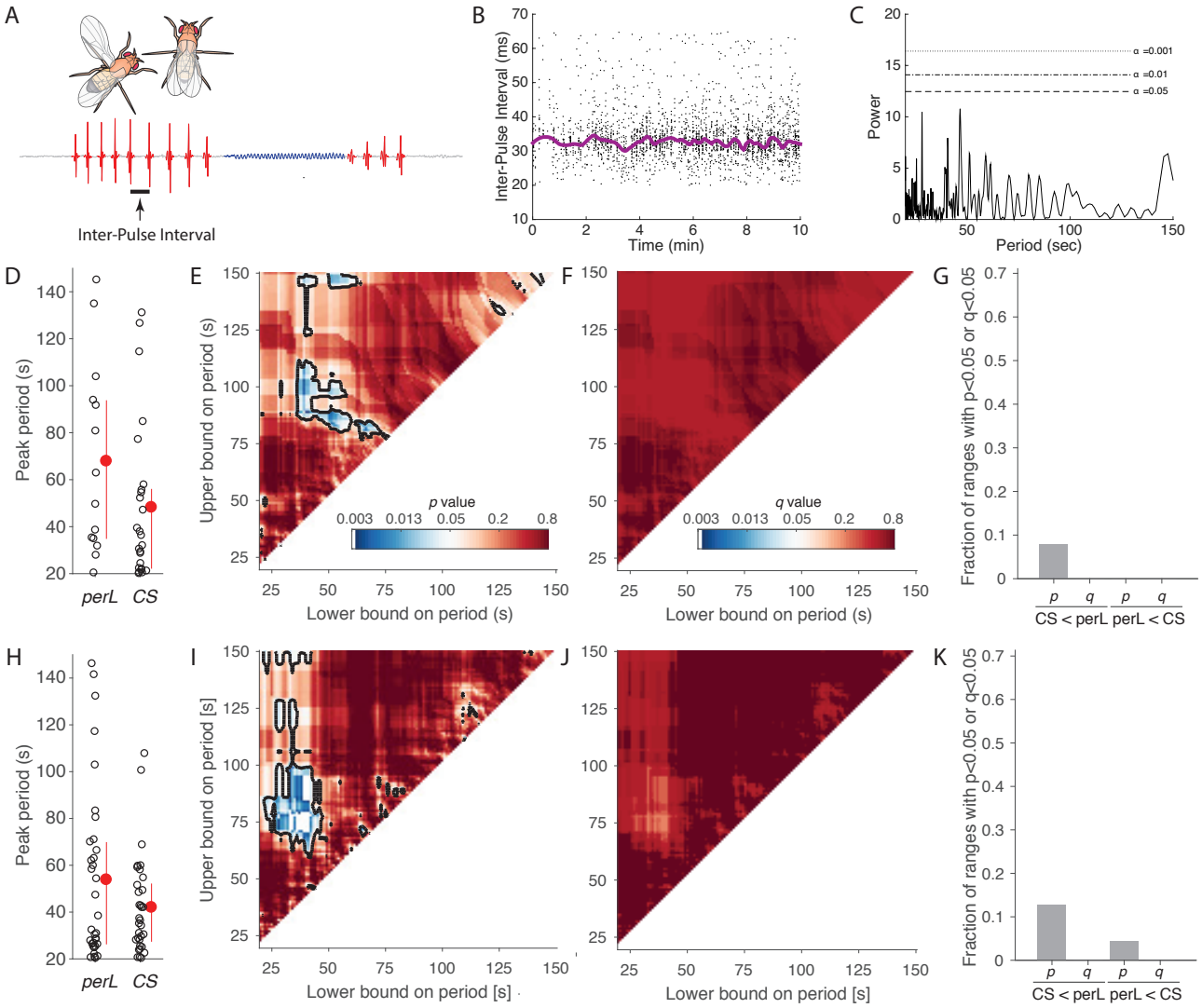


Figure 1. Genotype-specific periodicity cannot be detected in *Drosophila* courtship song. (A) *Drosophila* males produce courtship song, composed of pulses (red) and sines(blue), by extending and vibrating a wing. The inter-pulse interval is the time between consecutive pulses within a single train of pulses. (B) The inter-pulse interval varies widely during an entire bout of courtship. Within this wide range, the average inter-pulse interval often varies with an amplitude of several msec (purple line shows rloess fit with sliding window of 200 samples). (C) Lomb-Scargle periodogram analysis plotted for the range of 20 -150 sec of the inter-pulse interval data from panel (B). None of the peaks are significant at $p < 0.05$. (D) Comparison of the peak power between 20-150 sec from the Lomb-Scargle periodograms for the song data for the genotypes *period*^{L} (perL) and *Canton-S* (CS) manually-annotated by Kyriacou et al. (3). (Right-tailed T-test $p = 0.06$. Rank Sum $p = 0.10$.) (E) P values for period windows with different lower and upper bounds identify some windows with a nominal P value < 0.05. (F) False discovery rate q values for the windows shown in (E). None of these windows exhibit $q < 0.05$. (G) Fraction of ranges with significant (p or $q < 0.05$) for either the test of *Canton-S* less than *period*^{L} or *period*^{L} less than *Canton-S.* (H-K) Same as (D-G), except for newly collected song data from the same genotypes that was annotated using FlySongSegmenter.

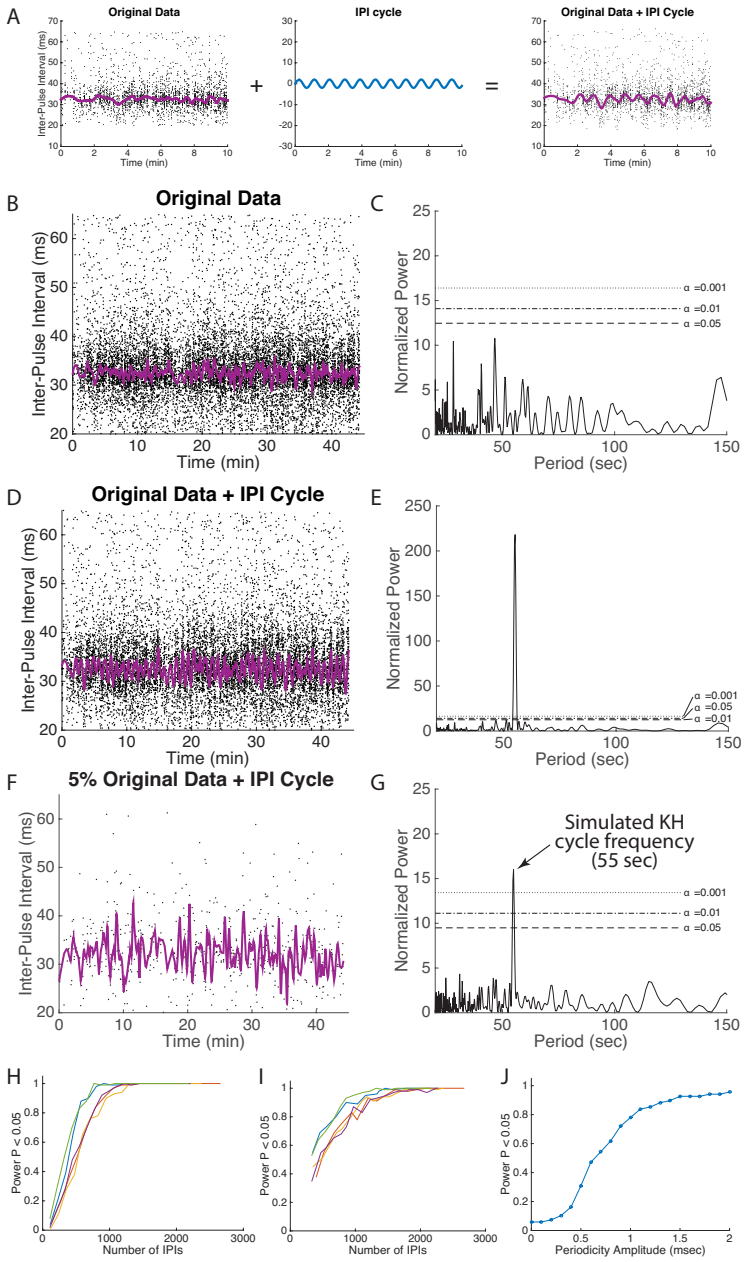


Figure 2. Simulations demonstrate that songs with at least 1000 inter-pulse interval measurements retain substantial power to detect rhythms, should they exist. (A) Schematic illustrating how a periodic cycle was added to raw inter-pulse interval data. Purple line illustrates the running mean of the raw data, on the left, and the data with an amplitude of 2 msec and a period of 55 sec imposed on the data, on the right. (B) One example of 45 minutes of inter-pulse interval data. Purple line shows that the running mean of data varies over time. (C) Lomb-Scargle periodogram of the data in (C) detects no significant rhythmicity. (D) The data from (B) with a 55 sec periodicity imposed. (E) The Lomb-Scargle periodogram of the data in (E) now reveals a highly significant peak at 55 sec, consistent with the simulated periodicity added to the raw data (F) Random removal of 95% of the inter-pulse interval data from (D). (G) Lomb-Scargle periodogram of the data in (F) still detects significant simulated periodicity, despite removal of most of the data. (H) Statistical power analysis of five songs (each song a different color) containing more than 10,000 inter-pulse interval events after simulated periodicity with 2ms amplitude was added to the raw data and individual inter-pulse 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 sec at $P < 0.05$. When songs retained at least 1000 inter-pulse interval events, the simulated periodicity could still be detected in more than 80% of the replicates. (I) The same analysis as in (H) except that ten-second bins of inter-pulse interval data were removed randomly. Songs with at least 1000 inter-pulse interval data remaining still contained significant periodicity in at least approximately 80% of the replicates. (J) 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 $P < 0.05$.

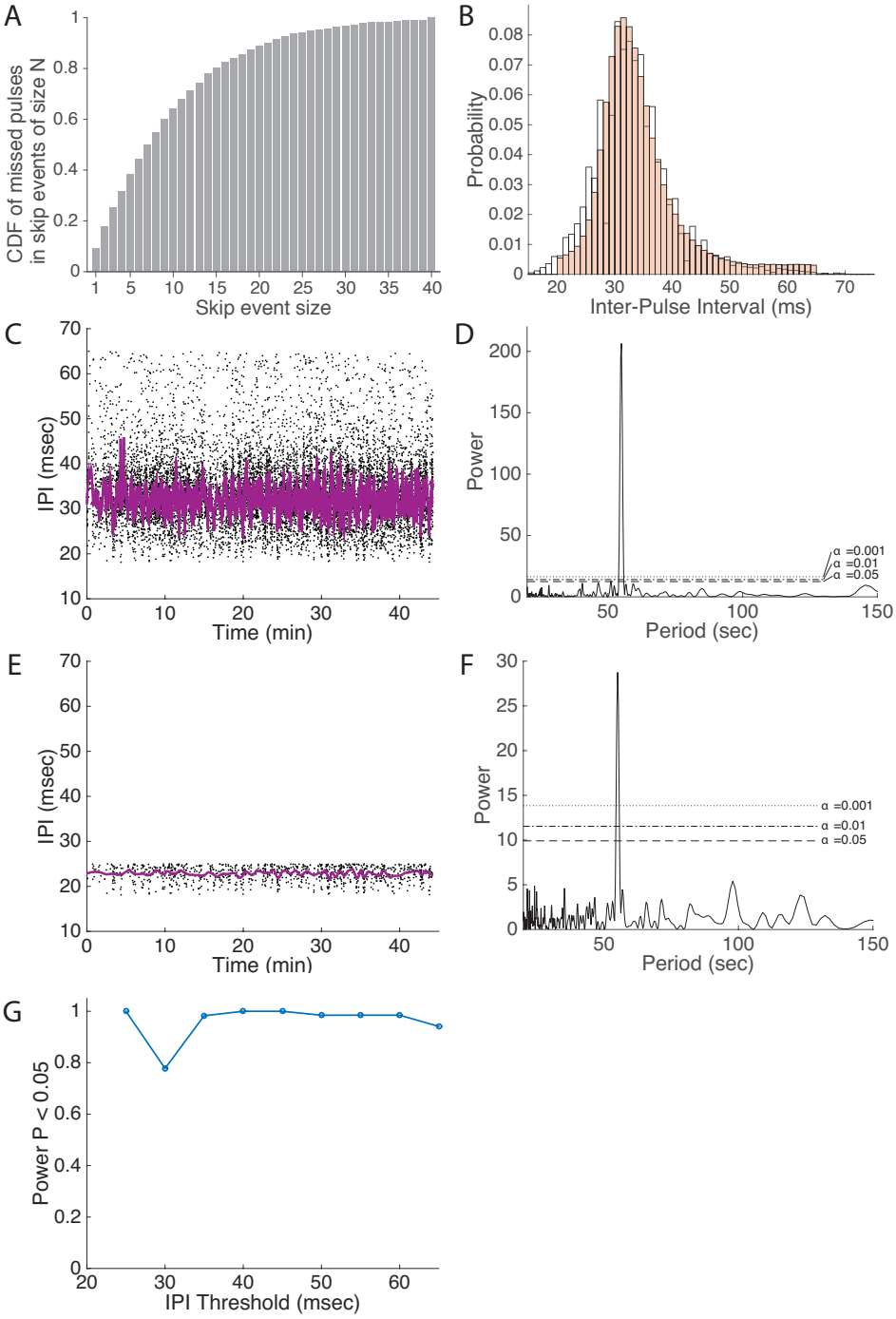


Figure 3. The specific inter-pulse interval threshold does not influence the statistical power to detect putative song rhythms. (A) Cumulative density function of the number of consecutively missed inter-pulse interval values in the data from Stern (2) illustrates that only 9% of missed pulses were singletons 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. (C) Example of one original song with 55 sec periodicity artificially imposed on the original inter-pulse interval data. (D) Lomb-Scargle periodogram of data in panel (C), revealing strong signal at 55 sec. (E) Same simulated data as in panel (C) with all inter-pulse interval values greater than 25 sec removed. (F) Lomb-Scargle periodogram reveals strong signal of the simulated periodicity at 55 sec, even though the data were thresholded at 25 sec. (G) Statistical power to detect simulated periodicity versus inter-pulse interval threshold for songs retaining at least 1000 inter-pulse interval values after thresholding. Periodicity between 50-60 sec was detected at $P < 0.05$ for most songs.

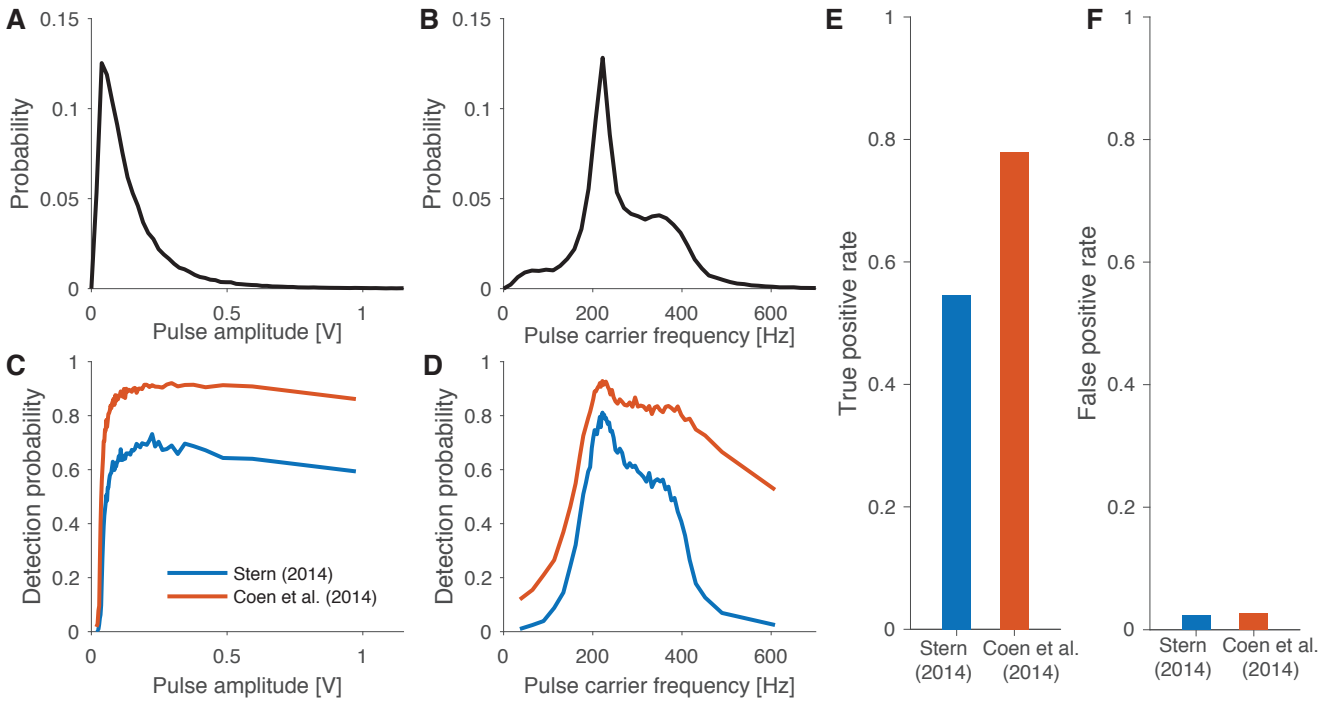


Figure 4. Modification of initialization parameters of FlySongSegmenter influences its performance in detecting pulses. (A, B) Distribution of pulse amplitudes (A) and carrier frequencies (B) for the pulses manually annotated in Kyriacou et al. (3). (C, D) Probability of detecting manually annotated 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).

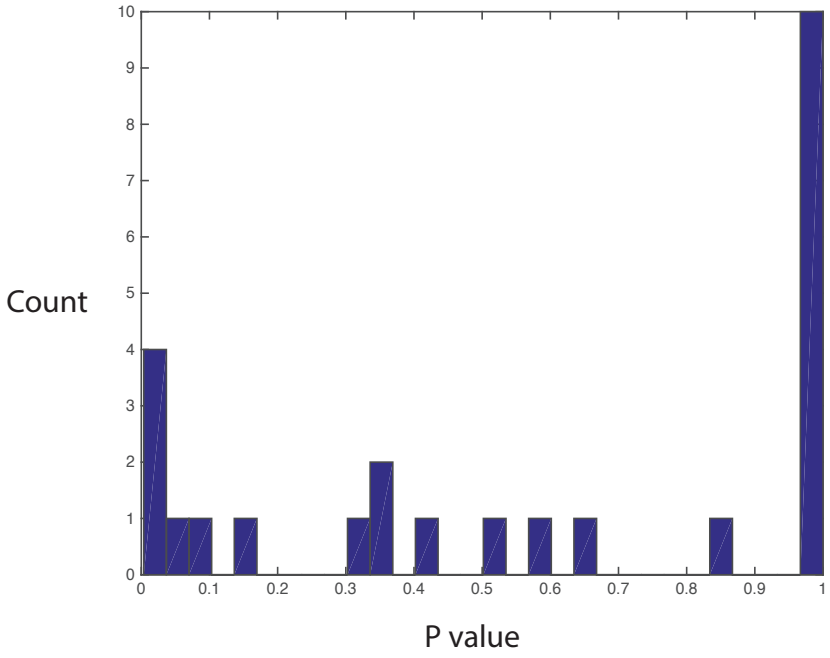


Figure S1. Distribution of p-values for the Lomb-Scargle periodogram peaks with maximum power between 20 and 150 sec for the *Canton-S* song data manually annotated by Kyriacou et al (3). Four of the peaks exhibit p-values < 0.05 and there is not an obvious excess of low p-values.

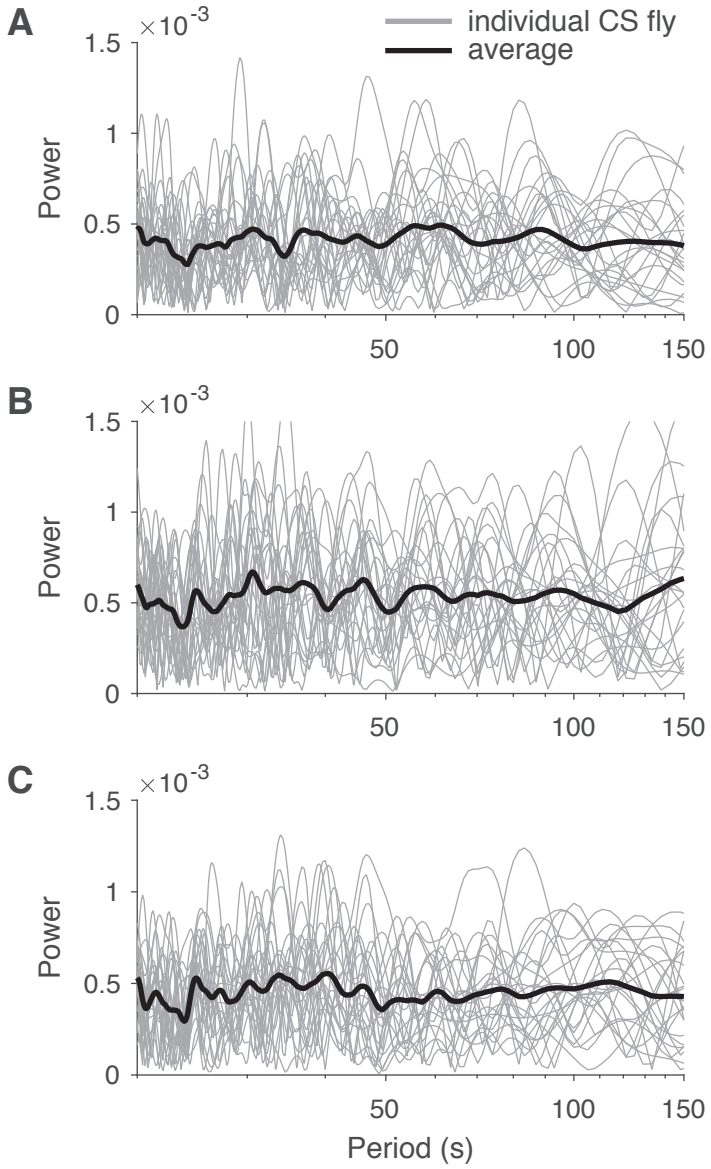


Figure S2. Lomb-Scargle periodograms for *Canton-S* song recordings. (A) From data manually annotated by Kyriacou et al. (3). (B) Automatically segmented data from Stern (3). (C) Automatically segmented data using segmentation parameters from Coen et al. (5). Individual recordings are shown in grey and average over all recordings is shown in black.

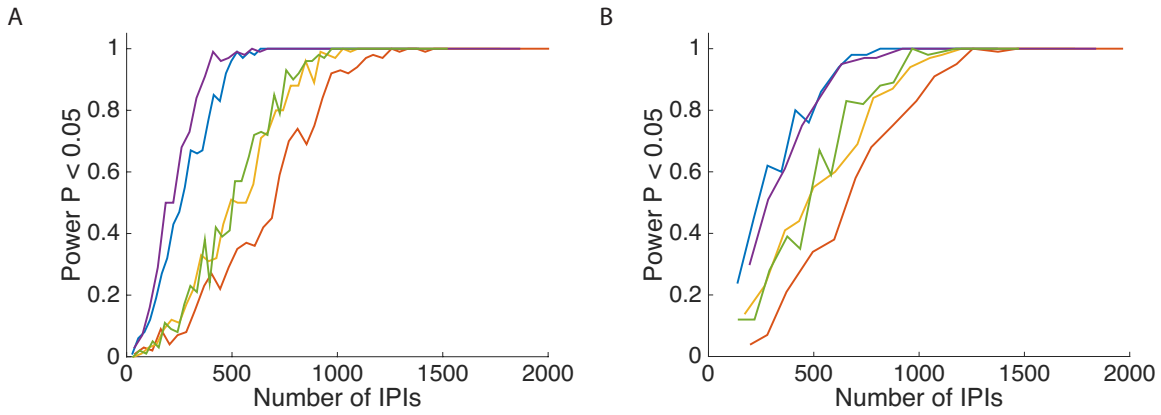


Figure S3. Statistical power analysis of short songs. (A, B) Simulated periodicity was added to five songs containing at least 10,000 inter-pulse interval (IPI) events in 45 minutes and then only the first 400 seconds of the song were analyzed. One hundred times, inter-pulse interval data were dropped either randomly (A) or 10 sec bins were dropped randomly (B) and Lomb-Scargle periodogram analysis was performed. The plots show the proportion of times out of 100 that periodicity was found between 50-60 sec with $P < 0.05$. In all cases, simulated periodicity could be detected at least 80% of the time in songs with at least 1000 inter-pulse interval events.

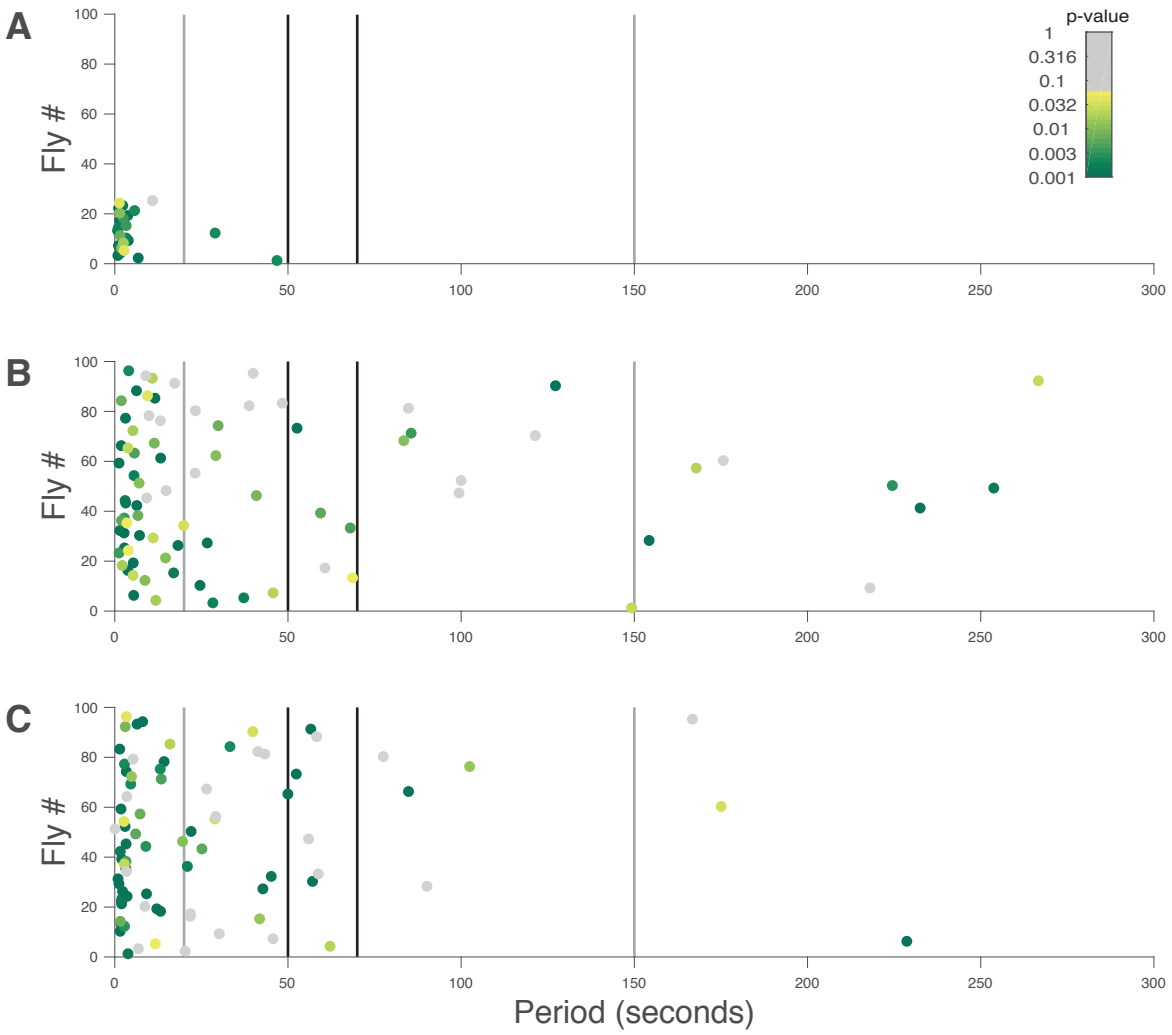


Figure S4. Period of maximum Lomb-Scargle periodogram peaks for inter-pulse interval measurements from multiple individual recordings. (A) Songs manually annotated by Kyriacou et al. (3). (B) Songs automatically segmented in Stern (2). (C) Songs from Stern (2) automatically segmented using parameters 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 are defined by the vertical black lines and the 20-150 sec range defined by Kyriacou et al. (3) are shown with vertical gray lines.

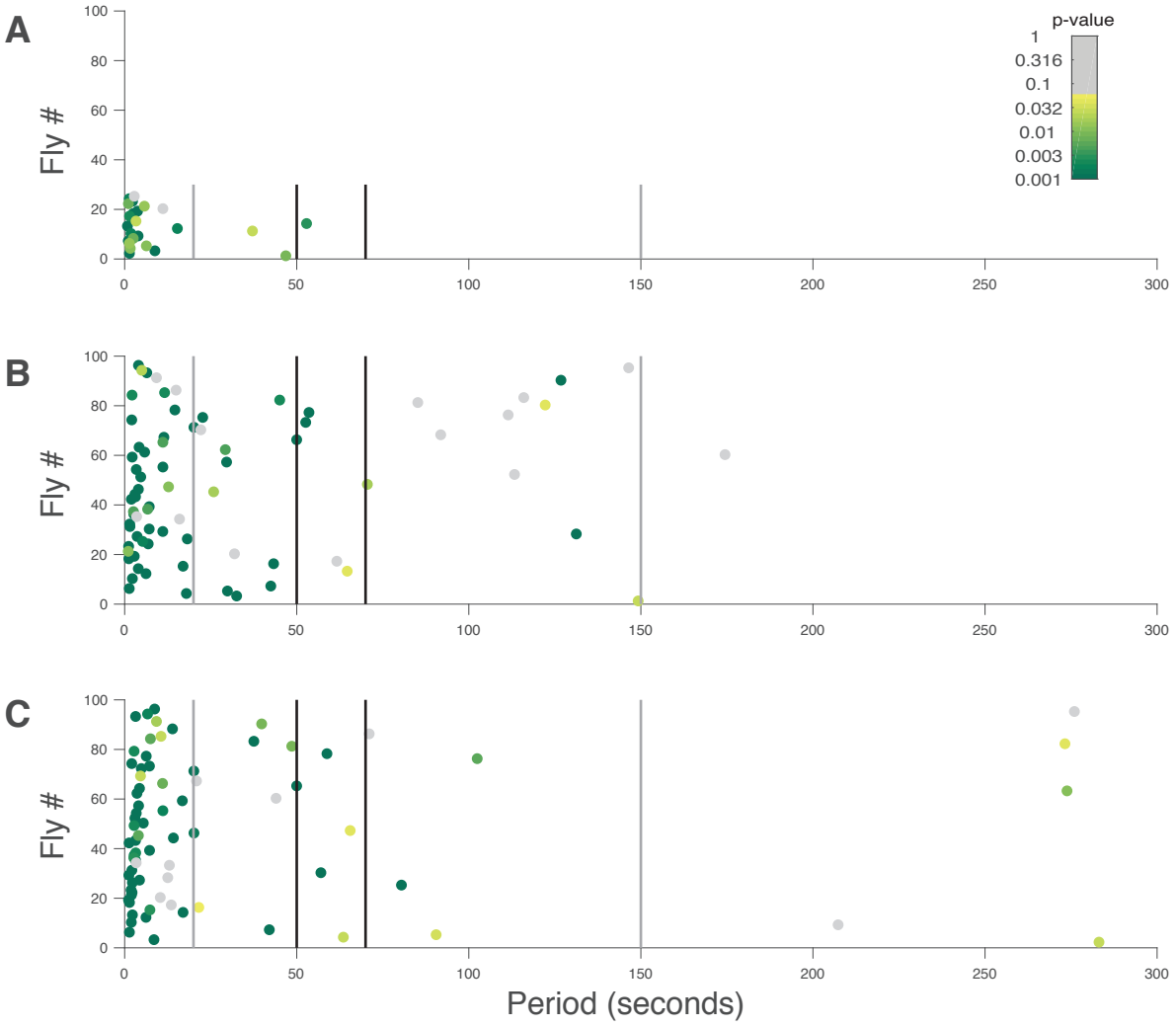


Figure S5. Period of maximum Lomb-Scargle periodogram peaks for inter-pulse interval measurements from multiple individual recordings with an inter-pulse interval cutoff of 55 msec. (A) Songs manually annotated by Kyriacou et al. (3). (B) Songs automatically segmented in Stern (2). (C) Songs from Stern (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 are defined by the vertical black lines and the 20-150 sec range defined by Kyriacou et al. (3) are shown with vertical gray lines.

Supplementary Material

Here we address several issues raised in Kyriacou et al. (1) that we did not have space to address in the main manuscript.

Inter-pulse interval cut-off and temperature control

Under the heading "Problem2: Inappropriate upper IPI cut-offs and poor temperature control," the authors 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.

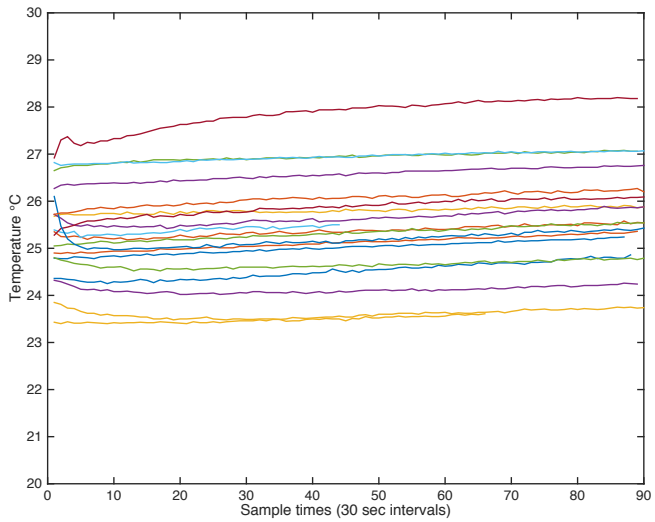
Inter-pulse interval cut-off: Kyriacou et al (1, 3) recommended that the IPI cut-off should scale with the mean inter-pulse interval for a genotype. They did not indicate precisely how the cut-off should scale with the mean. In their table S1, they indicated a "more appropriate cutoff" for each genotype 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	41.3	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	8.03	5.92	6.63	8.97	7.48	6.9	6.5

The mean inter-pulse interval varies by less than 10 ms, but the recommended cut-offs vary by 30 ms. The slope of the regression of mean inter-pulse interval and the recommended cut-off is 3.3 ($y = 3.3x – 47$). In essence, Kyriacou et al. assume that the standard deviation in inter-pulse interval increases 3.3 times faster than the mean inter-pulse interval. We find, in contrast, that the standard deviation in inter-pulse interval for each genotype is only weakly related to the mean IPI ($y = 0.17X + 0.72$ for automated data), suggesting that the standard deviation of the inter-pulse interval does not change 3.3 faster than the mean. Furthermore, in the main manuscript, we report simulations where we progressively reduced the cutoff for song with simulated rhythms. We find that the upper cut-off can be reduced 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 the cutoff of 65 ms influenced our ability to detect periodicity in the songs.

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) expressed concern that the experiments reported in Stern (2) had poor temperature control.

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.

Length of courtship

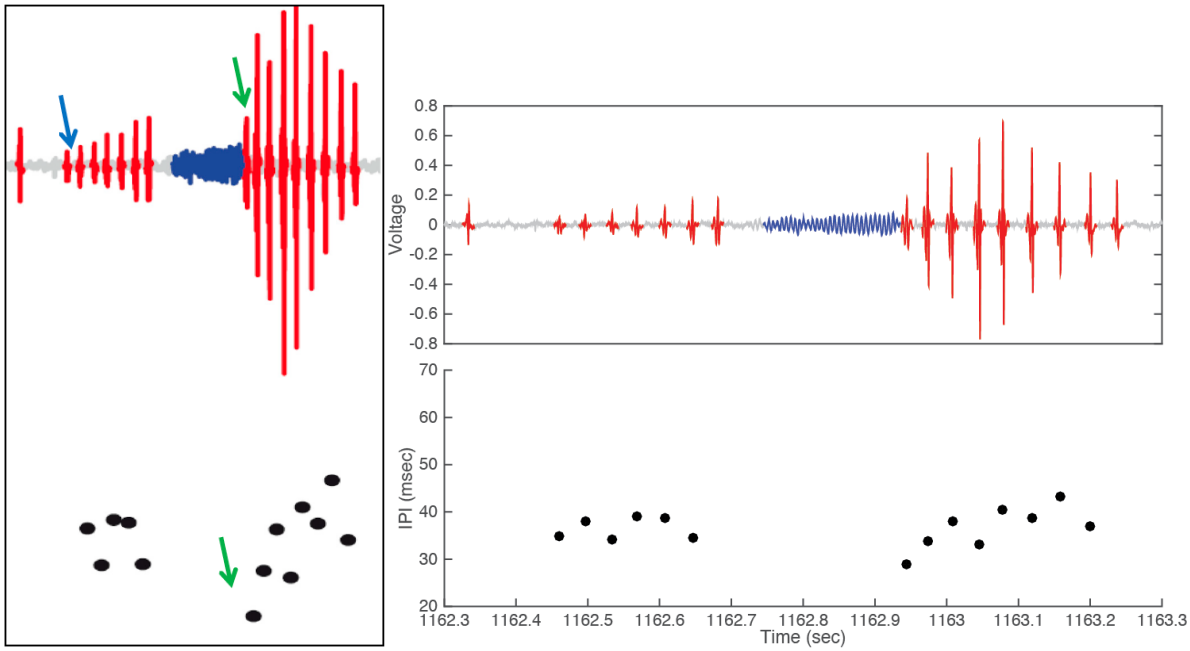
Under the heading "Problem 3: Unrealistic length of courtship," Kyriacou et al. (1) state that "courtship interactions under natural conditions are brief," lasting less than 30 sec and therefore question the use of 45 minute recordings of song. (Of course, if courtship really lasted less than 30 sec, then 50-60 sec periodicity could not be detected.) The key reference the authors cite for natural courtships (4) 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 (5). This paper 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.

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.



Supplementary Material

Here we address several issues raised in Kyriacou et al. (1) that we did not have space to address in the main manuscript.

Inter-pulse interval cut-off and temperature control

Under the heading "Problem2: Inappropriate upper IPI cut-offs and poor temperature control," the authors 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.

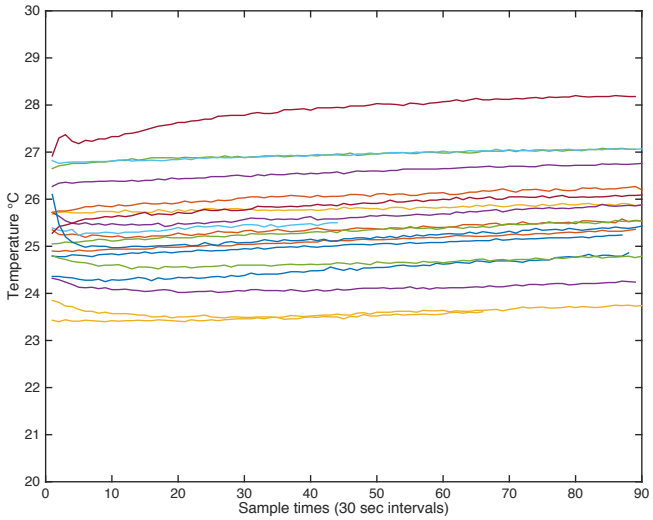
Inter-pulse interval cut-off: Kyriacou et al (1, 3) recommended that the IPI cut-off should scale with the mean inter-pulse interval for a genotype. They did not indicate precisely how the cut-off should scale with the mean. In their table S1, they indicated a "more appropriate cutoff" for each genotype 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 ms, but the recommended cut-offs vary by 30 ms. 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 3.1 times faster than the mean inter-pulse interval. We find, in contrast, that the standard deviation in inter-pulse interval for each genotype is only weakly related to the mean IPI ($y = 0.14x + 1.8$ for automated data), suggesting that the standard deviation of the inter-pulse interval does not change 3.3 faster than the mean. Furthermore, in the main manuscript, we report simulations where we progressively reduced the cutoff for song with simulated rhythms. We find that the upper cut-off can be reduced 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 the cutoff of 65 ms influenced our ability to detect periodicity in the songs.

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) expressed concern that the experiments reported in Stern (2) had poor temperature control.

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.



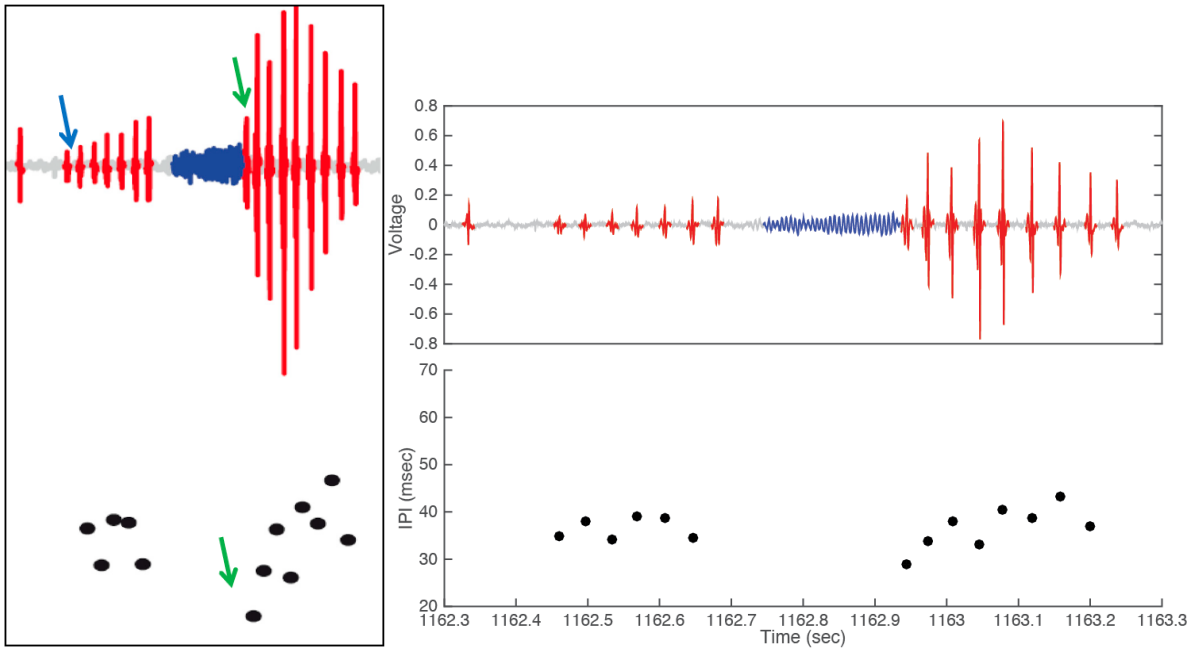
While these slight differences in temperature over the course of each experiment are expected to have a subtle effect on the inter-pulse interval, it is not clear that song periodicity should *disappear* as a result of these small temperature changes. One might imagine that the periodicity might differ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

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 (4) 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 (5). This paper 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.

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.