1   **Experimental and statistical re-evaluation provides no evidence for *Drosophila***

2   **courtship song rhythms**

3   Authors:

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

5   Arthur[1], and Mala Murthy[2,3]

6

7   Affiliations:

8   [1] Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA

9   [2] Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA

10   [3] Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

11   [4] University College London, Gower St, Bloomsbury, London WC1E 6BT, UK

12   [5] Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH 45229

13

14

15

16   Correspondence: sternd@janelia.hhmi.org

17

1

18   **Abstract**

19   From 1980 to 1992, a series of influential papers reported on the discovery, genetics, and

20   evolution of a periodic cycling of the interval between *Drosophila* male courtship song pulses.

21   The molecular mechanisms underlying this periodicity were never described. To reinitiate

22   investigation of this phenomenon, we performed automated segmentation of songs, but failed

23   to detect the proposed periodicity [Arthur BJ et al. (2013) *BMC Biol* 11:11; Stern DL (2014) *BMC*

24   *Biol* 12:38]. Kyriacou CP et al. [(2017) *PNAS* 114:1970-1975] report that we failed to detect song

25   rhythms because i) our flies did not sing enough and ii) our segmenter did not identify many of

26   the song pulses. Kyriacou et al. manually annotated a subset of our recordings and reported

27   that two strains displayed rhythms with genotype-specific periodicity, in agreement with their

28   original reports. We cannot replicate this finding and show that the manually-annotated data,

29   the original automatically segmented data, and a new data set provide no evidence for either

30   the existence of song rhythms or song periodicity differences between genotypes. Furthermore,

31   we have re-examined our methods and analysis and find that our automated segmentation

32   method was not biased to prevent detection of putative song periodicity. We conclude that

33   there is currently no evidence for the existence of *Drosophila* courtship song rhythms.

34

Significance statement

Previous studies have reported that male vinegar flies sing courtship songs with a periodic rhythm of approximately 55 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 did not detect all song events. 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 or in any dataset and that our original methods were not biased against detecting rhythms. We conclude that song rhythms cannot be detected.

\body
Introduction

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. 1a). 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, 2, 4–8). Like humans

during conversation, *Drosophila* males modulate their song based on sensory feedback from

their communication partner (4, 5).

Visual inspection of 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) and they reported

that the mean cycled with a periodicity of about 55 sec and was controlled, in part, by the

period gene, a gene required for circadian rhythms (11). Later papers demonstrated that

evolution of a short amino-acid sequence within the *period* protein caused species-specific

differences in this periodicity (11–14). These reports attracted considerable interest because

they implicated the *period* gene in ultradian rhythms, in addition to its well-known role in

circadian rhythms(15), and because it 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 had argued that the presence or detectability of the

rhythms was sensitive to assay conditions and methods of analysis (16). One of us, therefore,

4

69   replicated the methods of Kyriacou and Hall as closely as possible, but, again, song rhythms

70   could not be detected (2).

71   Kyriacou et al. (3) have recently questioned our previous conclusions. Here we focus on

72   three major assertions that they claim call our conclusions into doubt. First, we examine their

73   central claim that manual analysis of songs, but not automated analysis, reveals genotype-

74   specific song rhythms. We find that re-analysis of their manually-annotated data provides no

75   statistical support for genotype-specific rhythms. We also find no evidence for song rhythms in

76   the original dataset and a new larger dataset. Second, we examined their claim that the original

77   recordings contained insufficient data to detect rhythms and find that this claim is not

78   supported by simulation studies. Third, we examine their claim that the high false negative rate

79   of the automated song segmenter decreased the probability of detecting song rhythms and find

80   no evidence that the missing pulse events biased our analysis of song rhythms. Further, we

81   identify the major sources of false negative events in automated song analysis and illustrate

82   that minor modifications to initialization parameters substantially improve performance of the

83   song segmenter. Kyriacou et al. (3) also raised a number of minor concerns—such as how to

84   choose an appropriate inter-pulse interval cutoff, whether temperature was controlled

85   appropriately in our experiments, and whether songs produced beyond the first few minutes of

86   courtship should be analyzed—that we consider peripheral to the central questions raised and

87   therefore we have addressed these concerns (which are also unsupported by re-analysis) in the

88   SI Appendix.

89

90   **Results**

5

91   Earlier papers that identified song cycles employed several unusual methods of data

92   analysis that it is useful to review. First, continuous inter-pulse interval data were binned into

93   10 sec intervals. We reported previously that binning the data, together with the analysis of

94   relatively short songs, creates peaks in spectrogram analysis that fall within an artificially

95   narrowed frequency range, corresponding approximately to the frequency range originally

96   reported for the periodicity, and reduces the significance of periodiogram peaks (2 and see

97   below). Despite the fact that this procedure squeezes periodogram results into a narrow

98   frequency range, few songs contained peaks reaching a significance level of $p < 0.05$ (four of

99   149 songs, Fig. 3a of (2)), strongly suggesting that these peaks represent signals that cannot be

100   distinguished from noise. All of the previously reported "statistically significant" comparisons of

101   different genotypes are derived from analysis of mainly non-significant periodogram peaks. In

102   this re-evaluation, we do not discuss binning, but instead focus on other methodological issues.

103

104   **No evidence that manual song segmentation reveals genotype-specific song rhythms**

105

106   Kyriacou et al.'s (3) core finding is that different genotypes displayed different periodic

107   rhythms of the inter-pulse interval. This is also the most important discovery reported in earlier

108   papers on this subject (11–13, 17). Kyriacou et al. (3) manually annotated recordings made by

109   Stern (2) from a wild-type strain, *Canton-S*, and a strain carrying a *period* gene mutation, *per^L*,

110   for flies they categorized as singing "vigorously." We re-analyzed these data and the

111   automatically segmented data (2). Flies homozygous for *per^L* display circadian rhythms that are

112   longer than normal (15) and earlier papers have reported that *per^L* confers longer periods on

6

113   the inter-pulse interval rhythm (10–13). Kyriacou et al. (3) report a difference in the mean song

114   period between *Canton*-S and *per^L* with the manually annotated data, but not with the

115   automatically segmented data, suggesting that song cycles exist and display genotype-specific

116   frequencies and that the automatically segmented data is biased against detecting the song

117   rhythm.

118   Kyriacou et al. (3) used several methods to measure periodicity in the original time

119   series, which we discuss in more detail in the next paragraph. For approximately 85% of these

120   songs, these methods do not yield statistically significant signals in the frequency range of 20-

121   150 sec. Because most songs do not yield statistically significant peaks, Kyriacou et al. (3)

122   identified the peak with maximum power in the range of 20-150 sec for each song and

123   compared these values between genotypes. This is an unorthodox approach to data analysis. It

124   is equivalent to sampling outliers from a distribution of random noise and then performing

125   further statistics with these data. Nonetheless, Kyriacou et al (3) detected genotype-specific

126   song rhythms using this method and so, below, we accept this premise and investigate whether

127   there is statistical support for genotype-specific rhythms in the data. We start by examining

128   whether there is evidence for rhythms in individual songs.

129   The general model proposed for these song rhythms is that the inter-pulse interval

130   varies, on average, with a regular periodicity (10). Therefore, it should be possible to detect this

131   rhythmicity with appropriate methods of periodogram analysis. We have previously employed

132   Lomb-Scargle periodogram analysis (18–20) because this method does not require evenly

133   spaced samples and Kyriacou et al. (3) also adopted this method. For example, the Lomb-

134   Scargle periodogram of the time series in Fig. 1b is shown in Fig. 1c. In this case, despite the

7

135   obvious variation in inter-pulse interval values observed in Fig. 1b, there is no significant

136   periodicity between 20 and 150 sec. Kyriacou et al. (3) also employed Cosinor (21) and CLEAN

137   (22) for periodogram analysis. CLEAN does not produce a significance value for periodogram

138   peaks, so it is difficult to interpret. We find that Cosinor exhibits a high false positive rate (SI

139   Appendix, Fig. S1), and should be avoided for this type of analysis.

140   Kyriacou et al. (3) state that wild-type *D. melanogaster* songs exhibit periodicity

141   between 20 and 150 sec. Previously they reported that rhythms occurred with 50 – 60 sec

142   periodicity (10). Increasing the width of the periodicity window from 50-60 sec to 20-150 sec

143   increases the probability of detecting significant periods, but, even given this wide frequency

144   range, we observed that only 4 of the 25 manually annotated *Canton-S* songs and 3 of the 25

145   automatically segmented songs contained periodogram peaks that reached a significance level

146   of $P < 0.05$. (When we binned data in 10 sec bins, these values declined to 0 of 25 manually

147   annotated and 1 of 25 automatically segmented songs.) These significant peaks are not

148   localized to any particular narrow frequency range (SI Appendix, Fig. S1).

149   One reason to study non-significant peaks would be if periodicity is weak and not

150   detected reliably by periodogram analysis. This seems unlikely, since simulated song rhythms

151   can be detected with high confidence ((1, 2) and see below). Nonetheless, if periodogram

152   analysis is underpowered, then we expect to observe that the major peak in most songs should

153   display nearly-significant periodicity. In fact, we observe that 72% of p-values are greater than

154   0.2 (SI Appendix, Fig. S2). There is therefore no evidence that songs contain weak periodicity.

155   An alternative possible reason to include non-significant periodogram peaks in

156   downstream analysis is that the signal to noise of the periodicity is extremely low. An analogue

8

157 in neuroscience is that neural signals sometimes cannot be detected with high signal to noise

158 and that only by averaging over many trials of a stimulus presentation can a neural response be

159 detected robustly. We therefore examined the power distribution averaged over all the results

160 for each genotype. These plots are essentially flat, suggesting that there is no signal hidden in

161 the fluctuations of individual periodograms (SI Appendix, Fig. S3).

162 Given these observations, further analysis of these data seems unwarranted. However,

163 Kyriacou et al. (3) compared the maximum periodogram peaks between 20-150 sec for the

164 Canton-S and per^L recordings and found that the manually-annotated data showed a

165 statistically-significant difference in the mean period, although the automatically segmented

166 data did not (Fig. 3d of Kyriacou et al. (3)). This is the key result of their paper. We therefore

167 attempted to replicate this observation. For the manually annotated data from each song we

168 identified the peak in the periodogram of maximum power falling between a period of 20 and

169 150 sec. In contrast to their published results, we found that the average of the periods with

170 maximum power (most of which were not significant) was not significantly different at P < 0.05

171 between the genotypes Canton-S and per^L (Fig. 1d). We have no explanation for this

172 discrepancy between our statistical analysis and theirs.

173 Since there is no biological or quantitative justification for the particular frequency

174 ranges examined in any study, we wondered whether the results were sensitive to the

175 frequency range examined. We explored a wide range of possible frequency ranges and found

176 that the test statistic was sensitive to the precise frequency range selected (Fig 1e). Most

177 frequency windows do not generate a statistically significant difference between the genotypes

9

178   (Fig. 1e,g) and false discovery rate correction for multiple testing (23, 24) yields no frequency

179   ranges with significant results (Fig 1f,g).

180   Thus, there is no support for the specific results reported by Kyriacou et al. (3) and there

181   is no statistical support for defining song inter-pulse interval cycle periods as occurring within

182   any particular window. Most importantly, our analysis indicates that genotype-specific analysis

183   of non-significant periodogram peaks has no justification. It is difficult to reconstruct precisely

184   what steps in the analysis led previous reports to identify statistically significant genotype-

185   specific differences, but it is possible that previous studies may have serendipitously selected

186   frequency ranges that yielded significant results and/or did not properly control for multiple

187   testing.

188

189   **New data provide no evidence for genotypic specific song periodicities**

190

191   While we could not reproduce results reported by Kyriacou et al (3), we decided to take

192   their observation at face value as a preliminary result and to test directly whether genotype

193   specific song rhythms could be detected in a new, expanded data set. We recorded song from

194   33 *Canton-S* males and 34 *period^L* males. We identified the strongest periodogram peak in the

195   frequency range of 20-150 s for each song and found no significant difference between these

196   genotypes (Fig. 1h). We then compared test statistics across a wide set of frequency ranges, as

197   described above. We identified some frequency ranges that yielded significant results in the

198   predicted direction (Fig. 1i), with *period^L* rhythms slower than *Canton-S* rhythms, but for three

199   reasons we believe these results are spurious. First, and most importantly, none of these ranges

10

200   are significant after false discovery rate correction (Fig. 1j). Second, multiple frequency ranges

201   support the *opposite* conclusion, that *Canton-S* rhythms are slower than *period^L* rhythms (Fig.

202   1k). Third, the frequency ranges yielding significant comparisons only partially overlap with the

203   ranges found for the original dataset (c.f. Figs. 1e & 1i). In conclusion, there is not only no

204   evidence that song rhythms exist, there is also no evidence that reported genotype specific

205   differences in a song rhythm exist.

206   Putative song cycles cannot be identified in most automatically segmented song (2) and,

207   as we showed above, in most manually annotated song. In addition, when statistically

208   significant periodicity is detected, the frequencies of this periodicity do not cluster in a specific

209   frequency range, but instead are spread randomly across the entire frequency range examined

210   (SI Appendix, Fig. S5; Fig. 4 of Stern (2)). Finally, no songs are significant after correcting for

211   multiple comparisons (Fig. 1). All together, these results imply that the few *statistically*

212   significant periodicities that can be found do not carry *biological* significance.

213

214   **No evidence that low-intensity courtship provided insufficient data to detect song rhythms**

215

216   While we found no statistical evidence for the existence of song rhythms or of genotype

217   specific rhythms, we feel it is important to rebut several other statements made by Kyriacou et

218   al. (3). They state that rhythms can be detected only in songs produced by "vigorously" singing

219   males and write: "sporadic songs could not possibly provide any test for song cycles." It is not

220   clear if they mean that rhythms can be detected only in *songs* with many pulses or that only

221   *flies* that sing songs with many pulses ("vigorous singers") produce rhythms. Kyriacou et al. (3)

222   manually annotated songs from flies that they categorized as vigorous and we showed above

223   that significant periodicity can be found in only a minority of these songs and that these

224   significant values are not localized to a particular frequency range (SI Appendix, Fig. S1d).

225   Therefore, it is unlikely that only *flies* that sing songs with many pulses produce periodicity. We

226   therefore performed simulations to determine whether rhythms can be detected only in *songs*

227   with many pulses.

228   We previously investigated songs from 45-minute courtship recordings that contained at

229   least 1000 inter-pulse interval measurements (2). Kyriacou et al. (3) argued that more than 180

230   inter-pulse interval measurements per minute (or approximately 5000 events in a 45-minute

231   recording) should be identified to allow identification of song rhythms. To examine this claim,

232   we performed a statistical power analysis using songs with variable numbers of inter-pulse

233   interval measurements, where statistical power corresponds to the proportion of times

234   periodicity is detected in songs where periodicity has been artificially imposed on song data

235   (Fig. 2). We started with six 45-minute recordings of *Canton-S* from Stern (2) that contained

236   more than 10,000 inter-pulse interval measurements. None of these six songs yielded

237   statistically significant power in the frequency range between 50 and 60 sec (the range

238   originally defined to contain rhythms (10)) and one song produced a marginally significant peak

239   at 31.7 sec (P = 0.04), which falls between 20 and 150 sec (the range used by Kyriacou et al. (3)).

240   Figure 2d and 2e illustrate the inter-pulse interval data and periodogram for one of these songs.

241   Therefore, these songs do not contain strong periodicity in the predicted range and can serve as

242   a template to examine the power of Lomb-Scargle periodogram analysis to detect simulated

243   rhythms imposed on these data.

12

244   The initial reports of periodic cycles in the inter-pulse interval reported rhythms with a

245   mean period of 55 sec and an amplitude of approximately 2 ms (10). Therefore, we imposed a

246   55 second rhythm with an amplitude of 2ms on the six songs containing more than 10,000

247   inter-pulse interval measurements (Fig. 2a-c). We detected the simulated 55 sec rhythm in all

248   six songs with P-values < 10e-74 (example shown in Fig. 2f,g). We then randomly removed data

249   points from the songs iteratively and calculated the fraction of times we could detect the

250   simulated rhythm with P < 0.05. We removed data randomly from the dataset to simulate the

251   effect of failing to detect individual events in the song and we also removed chunks of data (in

252   10 sec bins) to simulate large gaps between song bursts, such as might be generated during

253   low-intensity courtship. We found that in both scenarios we could randomly remove at least

254   90% of the data and still detect simulated rhythms at least 80% of the time (example shown in

255   Fig. 2h,i; summary statistics shown in Fig. 2j and SI Appendix, Fig. S4a). That is, as long as songs

256   contained at least 1000 inter-pulse interval measurements, Lomb-Scargle periodogram analysis

257   detected simulated rhythms with power greater than 0.8. Similar results were found when we

258   analyzed only the first 400 sec of songs (SI Appendix, Fig. S4c,d). Furthermore, periodicity could

259   be detected with power greater than 0.8 when the amplitude of simulated periodicity was

260   greater than at least 1 msec (SI Appendix, Fig. 4b). These results were robust to noise in the

261   original periodicity. Song with a signal to noise ratio of as low as 0.25 could be detected with

262   power > 0.7 with sample sizes of at least 1000 inter-pulse interval measurements (Fig. 2k).

263   Similarly, periodicity could be detected reliably when we simulated a non-sinusoidal rhythm (SI

264   Appendix, Fig. Fig. S4e) and when periodicity was imposed for only a fraction of the total song

13

265   (SI Appendix). Thus, Lomb-Scargle periodogram analysis is a sensitive method for detecting

266   simulated periodicity, even in the presence of noise or discontinuities in the waveform.

267   Songs containing at least 1000 inter-pulse intervals provide sufficient data to identify

268   putative song cycles. In fact, we find that songs can be deeply corrupted by the absence of large

269   segments of song and simulated periodicity can still be detected.

270

271   **No evidence that the automated fly song segmenter biased the results**

272

273   Kyriacou et al. (3) expressed concern that our automated fly song segmenter displayed a

274   low true positive rate (the segmenter failed to detect approximately 50% of the pulses

275   identified through manual annotation) and produced some false positive calls (approximately

276   4% of events scored as pulses by the automated segmenter appear to be noise). They suggest

277   that these incorrect pulse event assignments could bias estimation of the mean inter-pulse

278   interval and therefore decrease the signal-to-noise of the periodic cycle, making it difficult to

279   detect a periodic signal. In principle, a large sample of incorrect calls could bias results, so we

280   investigated whether this was the case for our prior analyses. We used Kyriacou et al.'s (3)

281   manually-annotated dataset to investigate the potential for bias and to evaluate performance

282   of the automated segmenter.

283   When a single pulse event is not detected, the inter-pulse interval is then calculated as

284   the sum of the two neighboring real intervals. On average, this is approximately double the

285   average inter-pulse interval. The average inter-pulse interval for the *Canton-S* recordings

286   reported in Stern (2) is approximately 35 msec with a standard deviation of approximately 7

14

287   msec. Therefore, skipping a single pulse event is expected to result in inter-pulse interval

288   measurements of approximately 70 msec, but with considerable variance. Following Kyriacou

289   and Hall (16), Stern (2) employed a heuristic threshold of 65 msec to reduce the number of

290   spurious inter-pulse interval values. Therefore, in the specific case when a single pulse in a

291   train is missed, approximately one third of the incorrectly scored doublet inter-pulse interval

292   measurements would be shorter than 65 msec and are expected to contaminate the original

293   dataset.

294   However, this scenario applies only when one undetected pulse is flanked by two pulses

295   that are detected. Skipping more than one pulse would always result in inter-pulse interval

296   measurements that are excluded by the 65 ms threshold. We found, however, that only 9% of

297   the pulses missed by automated segmentation were singletons (SI Appendix, Fig. S6a). These

298   incorrect inter-pulse intervals contribute to a slight excess of inter-pulse intervals with high

299   values (SI Appendix, Fig. S6b). Lowering the inter-pulse interval threshold would, therefore,

300   remove most or all spurious inter-pulse intervals. Since our power analysis, discussed above,

301   revealed that periodogram analysis was robust to random removal of inter-pulse interval

302   events, as long as songs still contained at least 1000 values, loss of a small number of inter-

303   pulse intervals is not expected to hamper detection of rhythms. After reducing the inter-pulse

304   interval threshold to 55 msec, we still found no compelling evidence for significant periodicity in

305   the original data (SI Appendix, Fig. S7). Therefore, we explored the effect of reducing the inter-

306   pulse interval cutoff even further. In this case, we used all 68 *Canton-S* songs from Stern (2) and

307   retained for analysis only those songs that contained at least 1000 inter-pulse interval

308   measurements after imposing the new inter-pulse interval threshold. We explored a range of

15

309   cutoff values from 25 to 65 msec. We found that we could detect the simulated rhythm in most

310   songs with at least 1000 inter-pulse interval measurements remaining after thresholding, even

311   when the threshold was as low as 25 msec (Fig. 3). Therefore, we can find no evidence that

312   pulses missed by the automated song segmenter or the specific inter-pulse interval threshold

313   used in Stern (2) prevented detection of song rhythms.

314   Although detection of putative song rhythms is robust to dropped pulses in songs that

315   retain at least approximately 1000 inter-pulse intervals, it is worth reviewing briefly why the

316   segmenter failed to detect certain pulses in recordings reported in Stern (2). The first step of

317   song segmentation involves detection of pulse-like signals and sine-like signals (1). In

318   subsequent steps, the segmenter filters out many kinds of sounds that were originally classified

319   as song pulses. Both the initial detection of pulses and subsequent filtering steps are sensitive

320   to multiple parameters. These parameters are specified prior to segmentation and can be

321   modified to enhance performance of the segmenter for different recordings. We identified two

322   primary causes for missed pulses. First, Stern (2) recorded song in larger chambers than those

323   used previously with these microphones (1), to match the chamber size used by Kyriacou & Hall

324   (10). This larger chamber with one microphone had reduced sensitivity compared to the

325   original smaller chamber. The segmenter thus tended to miss pulses of lower amplitude, which

326   are hard to automatically differentiate from noise, and this explains approximately 35% of the

327   missed pulses (SI Appendix, Fig. S8a, c).

328   The second major cause of missed pulses is that *Drosophila* males produce pulses with a

329   range of carrier frequencies (tones). The higher frequency pulses tend to resemble other non-

330   song noises, like grooming, and a user can set parameters in the segmenter to attempt to

16

331   exclude these non-song noises based on the carrier frequency of the event. Stern (2) used

332   parameters to minimize the false positive rate, including a relatively low carrier frequency

333   cutoff for pulses. The lower pulse frequency threshold used by Stern (2) explains approximately

334   42% of the missed pulses (SI Appendix, Fig. S8b,d). Using the same software with different

335   parameters (from Coen et al. (5)) recovers many of these high-frequency pulses without

336   substantially increasing the false positive rate (SI Appendix, Fig. S8c-f).

337   Above, we showed that including more pulse events, by manual annotation, did not

338   increase the probability of detecting song rhythms. Therefore, there is no evidence that the

339   data resulting from the song segmenter parameters used in Stern (2) generated a data set that

340   was biased against detection of song rhythms. While the song segmenter does not detect all

341   pulse events that can be detected by manual annotation, the segmenter does provide data sets

342   that are several orders of magnitude larger than those that can be generated by manual

343   annotation, which has allowed discovery of multiple new phenomena related to *Drosophila*

344   courtship song (4–6). In addition, the sensitivity of the song segmenter can be improved with

345   optimization of initial parameters, as expected of any segmentation algorithm.

346

347   **Discussion**

348   We cannot detect a periodic cycling of the inter-pulse interval in *Drosophila* courtship

349   song even in the songs manually annotated by Kyriacou et al. (3) and used as evidence for

350   periodicity in their paper. While it is impossible to prove a negative, our results agree with

351   previous analyses that have concluded that there is no statistical evidence that these rhythms

352   exist (1, 2). In particular, by exploring some of the relevant parameter space with statistical

17

353   tests on the song that was manually-annotated by Kyriacou et al. (3), we find that subsets of

354   parameters sometimes produce p-values lower than 0.05, but that (1) few regions of parameter

355   space generate "significant" results, (2) these "significant" regions are scattered apparently

356   randomly in parameter space, and (3) none of these "significant" results survive multiple test

357   correction (Fig. 1).

358   Previously, we offered one explanation for how apparent song rhythms may have been

359   detected. We found that binning data from short songs confined the periodogram peaks with

360   maximum power close to the range reported as the song cycle (2). While few of these peaks

361   reached statistical significance, previous authors have accepted these peaks as "signal" and

362   performed statistical analyses to compare the peaks between genotypes. All "statistically

363   significant" results from earlier papers were derived mainly from non-significant peaks in

364   periodogram analysis and from relatively small sample sizes (usually fewer than 10 flies of each

365   genotype), so it is questionable whether these derivative statistics are valid. Genotype-specific

366   periodicities reported in earlier papers may have resulted, by chance, from studies of a small

367   number of short songs that fortuitously led to occasional apparent replication of the original

368   observations.

369   There may be a more prosaic explanation for the initial discovery of song cycles. Every

370   fly produces highly variable inter-pulse intervals. In addition, a running average of these data

371   reveals that the average inter-pulse interval cycles up and down (Fig. 1b), similar to the

372   temporally-binned data first reported by Kyriacou and Hall (10). There is no debate about this

373   observation. The claim in dispute is that the average inter-pulse interval cycles regularly. We

374   can find no evidence for this claim. It is easy to imagine, however, that visual examination of

18

375   short recordings of song would make it appear as if the mean inter-pulse interval cycled

376   regularly.

377   The extraordinary within-fly variation in the inter-pulse interval and in the mean inter-

378   pulse interval may result from multiple causes, including the possibility that male flies respond

379   to ever-changing cues during courtship and modulate their inter-pulse interval to optimize their

380   chances of mating. Individual *Drosophila* males modulate specific aspects of their courtship

381   song based on their own patterns of locomotion and in response to feedback from females,

382   including the transition between sine and pulse song (5) and the amplitude of pulse song (4).

383   There is additional evidence that males modulate the carrier frequency of sine song (1). We

384   hypothesize that male flies also modulate their inter-pulse interval in response to specific

385   internal or external cues.

386   We can find no statistical evidence for periodicity of the inter-pulse interval in individual

387   courtship songs and no evidence that comparisons of the strongest periodogram peaks from

388   each song identify genotype-specific rhythms. These results hold *both* for the songs manually

389   annotated by Kyriacou et al. (3) and for two independent large datasets automatically

390   annotated with FlySongSegmenter using optimized parameters. At this time, a conservative

391   assessment of the problem is that *Drosophila* courtship song rhythms and genotype-specific

392   effects on these rhythms cannot be replicated.

393

394   **Methods**

19

395   Courting fruit flies of Oregon-R and *per^L* were recorded as described previously (2). All analyses

396   were performed in Matlab. All data and code are freely available, as described in the Software

397   and Data Availability section. Further methods can be found in SI Appendix.

398

399   **Acknowledgements**

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

401

402   **Software and data availability**

403

404   Computer code for all analyses described in this paper is available at

405   https://github.com/murthylab/noIPIcycles. Code for the version of FlySongSegmenter used in

406   Cohen et al. (5) is available at https://github.com/murthylab/songSegmenter. The raw and

407   segmented song data for the new song recordings is available at

408   https://www.janelia.org/lab/stern-lab/tools-reagents-data.

409

410   **References**

411   1.   Arthur BJ, Sunayama-Morita T, Coen P, Murthy M, Stern DL (2013) Multi-channel

412   acoustic recording and automated analysis of Drosophila courtship songs. *BMC Biol*

413   11:11.

414   2.   Stern DL (2014) Reported Drosophila courtship song rhythms are artifacts of data

415   analysis. *BMC Biol*.

416   3.   Kyriacou CP, Green EW, Piffer A, Dowse HB, Takahashi JS (2017) Failure to reproduce

417   period-dependent song cycles in Drosophila is due to poor automated pulse-detection

418   and low-intensity courtship. *PNAS*. doi:10.1073/pnas.1615198114.

419 4. Coen P, Xie M, Clemens J, Murthy M (2016) Sensorimotor Transformations Underlying

420   Variability in Song Intensity during Drosophila Courtship. *Neuron* 89(3):629–644.

421 5. Coen P, et al. (2014) Dynamic sensory cues shape song structure in Drosophila. *Nature*.

422   doi:10.1038/nature13131.

423 6. Ding Y, Berrocal A, Morita T, Longden KD, Stern DL (2016) Natural courtship song

424   variation caused by an intronic retroelement in an ion channel gene. *Nature*

425   536(7616):329–332.

426 7. Shirangi TR, Wong AM, Truman JW, Stern DL (2016) Doublesex Regulates the

427   Connectivity of a Neural Circuit Controlling Drosophila Male Courtship Song. *Dev Cell*

428   37(6):533–544.

429 8. Shirangi TR, Stern DL, Truman JW (2013) Motor Control of Drosophila Courtship Song.

430   *Cell Rep* 5(3):678–686.

431 9. Bennet-Clark HCC, Ewing AW, Bennet-Clark HCC (1968) The courtship songs of

432   Drosophila. *Behaviour* 31(3):288–301.

433 10. Kyriacou CP, Hall JC (1980) Circadian rhythm mutations in Drosophila melanogaster affect

434   short-term fluctuations in the male's courtship song. *PNAS* 77(11):6729–6733.

435 11. Zehring WA, et al. (1984) P-element transformation with period locus DNA restores

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

437 12. Wheeler DA, et al. (1991) Molecular transfer of a species-specific behavior from

438   Drosophila simulans to Drosophila melanogaster. *Science* 251(4997):1082–5.

21

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. Refinetti R, Lissen GC, Halberg F (2013) *Procedures for numerical analysis of circadian rhythms* doi:10.1080/09291010600903692.Procedures.

22. Roberts DH, Lehár J, Dreher JW, Lehar J (1987) Time series analysis with CLEAN. I. Derivation of a spectrum. *Astron J* 93(4):968–989.

23. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and

461   powerful approach to multiple testing. *J R Stat Soc B* 57(1):289–300.

462   24.   Colquhoun D, London C (2014) An investigation of the false discovery rate and the

463   misinterpretation of P values. *R Soc Open Sci* 1:1–15.

464

465

466   Figure Legends

467

468   Figure 1. Genotype-specific periodicity cannot be detected in Drosophila courtship song. (A)

469   Drosophila males produce courtship song, composed of pulses (red) and sines (blue), by

470   extending and vibrating a wing. The inter-pulse interval is the time between consecutive pulses

471   within a single train of pulses. (B) The average inter-pulse interval varies over time. (Purple line

472   is rloess fit with sliding window of 200 samples). (C) Lomb-Scargle periodogram analysis of the

473   inter-pulse interval data from panel (B) plotted for the range of 20 -150 sec. None of the peaks

474   are significant at $p < 0.05$. (D) Comparison of the peak power between 20-150 sec from the

475   Lomb-Scargle periodograms for the song data for the genotypes periodL (perL) and Canton-S

476   (CS) manually-annotated by Kyriacou et al. (3). Red points and lines represent mean ± 1 SD for

477   each genotype. (Right-tailed T-test $p = 0.06$. Rank Sum $p = 0.10$.) (E) P-values for period

478   windows with different lower and upper bounds. (F) False discovery rate q values for the

479   windows shown in (E). (G) Fraction of ranges with significant (p or $q < 0.05$) for either the test of

480   Canton-S less than periodL or periodL less than Canton-S. (H-K) Same as (D-G) for newly

481   collected song data from the same genotypes annotated using FlySongSegmenter. (H) (Right-

482   tailed T-test $p = 0.06$. Rank Sum $p = 0.45$.)

483

484   Figure 2. Simulations to explore power to detect rhythms, should they exist. (A-C) Example of

485   how a periodic cycle was added to raw inter-pulse interval (IPI) data. Purple line in (A) illustrates

486   the running mean of the raw data. Blue line in (B) shows a periodic rhythm with an amplitude of

487   2 msec and a period of 55 sec. Original data with simulated periodicity is shown in (C). (D) One

24

488   example of 45 minutes of inter-pulse interval data. Purple line shows running mean. (E) Lomb-

489   Scargle periodogram of data in (D) does not detect periodicity. (F) Data from (D) with a 55 sec

490   periodicity imposed. (G) Lomb-Scargle periodogram of data in (F) now reveals a highly

491   significant peak at 55 sec, consistent with the simulated periodicity. (H) Random removal of

492   95% of the inter-pulse interval data from (F). (I) Lomb-Scargle periodogram of the data in (H)

493   detects significant periodicity. (J) Power analysis of six songs (each song a different color)

494   containing more than 10,000 inter-pulse interval events after 55 sec periodicity was added and

495   individual inter-pulse interval events were removed randomly. Power equals the fraction of

496   times out of 100 that a song contained a rhythm with significant periodicity between 50 and 60

497   sec at P < 0.05. (K) Power to detect simulated noisy periodicity versus number of IPIs remaining

498   after random removal of IPIs. Means of simulations for six songs containing more than 10,000

499   inter-pulse interval measurements are shown. Examples of simulated noisy rhythms are shown

500   to the right. Colorbar shows power to detect simulated rhythm.

501

502   Figure 3. The specific inter-pulse interval threshold does not influence the statistical power to

503   detect putative song rhythms. (A) Example of one original song with 55 sec periodicity

504   artificially imposed on the original inter-pulse interval data. (B) Lomb-Scargle periodogram of

505   data in panel (A), revealing strong signal at 55 sec. (C) Same simulated data as in panel (A) with

506   all inter-pulse interval values greater than 25 sec removed. (D) Lomb-Scargle periodogram

507   reveals strong signal of the simulated periodicity at 55 sec, even though the data were

508   thresholded at 25 sec. (E) Power to detect simulated periodicity versus inter-pulse interval

509   threshold for songs retaining at least 1000 inter-pulse interval values after thresholding.