

# Inconsistencies in *C. elegans* behavioural annotation

Balázs Szigeti<sup>1,2\*</sup>, Thomas Stone<sup>1</sup>, Barbara Webb<sup>3</sup>

<sup>1</sup> Neuroinformatics Doctoral Training Centre, University of Edinburgh, UK

<sup>2</sup> OpenWorm foundation, San Diego, CA, USA

<sup>3</sup> School of Informatics, University of Edinburgh, UK

\* Corresponding author: [b.szigeti@sms.ed.ac.uk](mailto:b.szigeti@sms.ed.ac.uk)

## Abstract

High quality behavioural annotation is a key component to link genes to behaviour, yet relatively little attention has been paid to check the consistency of various automated methods and expert judgement. In this paper we investigate the consistency of annotation for the ‘Omega turn’ of *C. elegans*, which is a frequently used behavioural assay for this animal. First the output of four Omega detection algorithms are examined for the same data set, and shown to have relative low consistency, with F-scores around 0.5. Consistency of expert annotation is then analysed, based on an online survey combining two methods: participants judged a fixed set of predetermined clips; and an adaptive psychophysical procedure was used to estimate individual’s threshold for Omega turn detection. This survey also revealed a substantial lack of consistency in decisions and thresholds. Such inconsistency makes cross-publication comparison difficult and raises issues of reproducibility.

## 1 Introduction

Traditionally, behavioural annotation has been done manually, with the known weakness of inherent variability, as well as being labour intensive. In the current era of big data biology, there is an increasing tendency for behavioural annotation to be automated [1, 2]. Automated methods can obviously scale to significantly larger data sets, but they are also supposed to improve consistency by removing human judgement from the process. However, the self-consistency of automated methods does not guarantee consistency between different methods. Furthermore, these algorithms are typically validated relative to a human produced ‘ground truth’ dataset [3–7]. This evaluation process raises the possibility that algorithms are trained to learn the same observational biases - and variance - that are inherent to human annotation. Given that different research groups often use different

31 annotation methods, a lack of consistency in their output could make comparison  
32 of published results from these groups difficult.

33

34 In this paper we specifically address the consistency of the behavioural anno-  
35 tation of the nematode worm *Caenorhabditis elegans* (*C. Elegans*), focusing on a  
36 particular worm behaviour, the Omega turn. Omega turns occur during reorienta-  
37 tions, with the animal adopting a shape resembling the Greek letter  $\Omega$ , see Figure  
38 1A for a representative example. This behaviour was chosen as it is often treated  
39 as a discrete, well defined element of worm behaviour [5, 7–10].

40

41 Our Omega turn consistency check has two components. First we examine the  
42 consistency of four Omega detection algorithms from the literature [4–7]. Second,  
43 we present the results of an on-line survey where we have invited experts to score  
44 Omega turns. The survey itself had two underlying components. Participants  
45 scored a set of predetermined clips and we have also employed an adaptive psy-  
46 chophysical method to identify individual’s threshold for Omega turns.

47

48 The results show that both expert annotation and algorithms are surprisingly  
49 inconsistent, and greater effort may be needed to ensure annotation methods pro-  
50 vide a reliable basis for studies that include behavioural assays.

## 51 2 Methods

### 52 2.1 Behavioural data

53 This study used data from the *C. elegans* behavioural database (CBD) [5]. The  
54 database consists of worm videos and corresponding feature files that contain a  
55 number of precalculated feature time series (such as speed, eccentricity, eigen-  
56 worm coefficients, etc.). We examined 776 experiments, all with hermaphrodite  
57 N2 worms. Worms were placed on a plate covered with a bacterial layer and the  
58 behaviour was recorded after a 30 minute habituation period. Each video is ap-  
59 proximately 15 minutes long, so in total 194 hours of worm behaviour was analysed.

60

61 During Omega turns, the worm can contact itself, producing an intersecting  
62 shape in the videos, and for these frames it is difficult to extract a biologically  
63 meaningful skeleton [6, 11]. As a consequence these ‘coiling’ frames are not pro-  
64 cessed in the CBD and the features for the corresponding frames are not calculated.  
65 If the resulting gap in the video was smaller than 20 consecutive frames (0.6 sec)  
66 then linear interpolation was used to gain a proxy for the features. This interpo-  
67 lation method is not reliable for longer gaps, hence Omega events that contained

68 longer gaps were discarded.

## 69 **2.2 Consistency of Omega turn detection algorithms**

### 70 **Algorithms**

71 Four algorithms have been taken from the literature to examine their consistency  
72 with each other. The algorithms are from the Zentracker package [4], the *C. elegans*  
73 behavioural database (CBD) [5], a computer vision based study to detect such  
74 events [6] and from a recent publication studying search behaviour [7]. Common  
75 to all these methods is that they detect Omega turns if a feature or a combination  
76 of features exceeds a user defined threshold. For example, [5] uses the midbody  
77 bend as the defining property of Omega turns. Note that this is not an exhaustive  
78 list of Omega turn detection algorithms. These particular algorithms have been  
79 chosen because the code used for the original publication was readily available.

### 80 **Consistency quantification**

81 To summarise annotation consistency we report the precision (positive predictive  
82 value) and sensitivity (also known as recall and true positive rate) [12]. Precision  
83 is the ratio of true positive events to all events recognised, while sensitivity is  
84 the proportion of true positives to all reference events. Mathematically they are  
85 expressed as

$$86 \quad Precision = \frac{TP}{TP + FP}, \quad Sensitivity = \frac{TP}{TP + FN}, \quad (1)$$

86

87

88 where  $TP$ ,  $FP$  and  $FN$  are true positive, false positive and false negative  
89 respectively. For example, if one algorithm is taken as the reference for Omega  
90 events, a true positive occurs for the comparator algorithm when it selects the  
91 same event (a  $TP$  was counted if at least 50% of the frames identified as part of  
92 an Omega turn overlapped); a false positive when it selects an event not labelled  
93 by the reference algorithm; and a false negative when it fails to select an event  
94 that was labelled by the reference algorithm. Precision and sensitivity are often  
95 combined to a single number summary, the F-score, which is defined as:

$$F = \frac{2(Precision \times Sensitivity)}{Precision + Sensitivity}. \quad (2)$$

## 96 **Threshold tuning**

97 The consistency between annotation algorithms is likely to be affected by pa-  
98 rameter settings. Therefore we calculated the results first with the original feature  
99 thresholds (taken from the publication) for each method, and then with the thresh-  
100 olds altered so as to find the best match between each pair of algorithms that could  
101 be obtained by parameter adjustment.

102  
103 To find the best match, each algorithm was run 25 times with different thresh-  
104 olds. For each run the difference in the threshold was increased or decreased by  
105 2.5% of the initial value. Therefore a range 70%-130% of the initial threshold val-  
106 ues were scanned. Lower percentages correspond to a more permissive definition  
107 (i.e. more events classified as Omega turns), but some scales had to be inverted.  
108 For example [4]’s method uses an upper bound on ‘eccentricity’ and a lower bound  
109 on ‘solidity’. Therefore to make the run associated with 70% more permissive, the  
110 eccentricity scale had to be inverted.

## 111 **2.3 Community survey of Omega turns**

### 112 **Survey structure**

113 To compare the consistency of expert Omega turn detection an online survey was  
114 developed <sup>1</sup>. After a brief registration, participants were shown 40 short (2-5s)  
115 clips of Omega events and were asked to indicate, using a button press, if each  
116 was an Omega turn or not. Participants were also asked to rate their confidence  
117 to detect Omega turns on a scale 1-5 (with 5 being very confident).

118  
119 In the survey we wanted to include ambiguous, wide amplitude turns that one  
120 may or may not consider an Omega turn. Therefore to select events for the survey  
121 we have run the Omega detection algorithm by [6] on the CBD videos, but with  
122 the threshold reduced to 75% of its original value. Using this criteria 1526 Omega  
123 like events were detected.

124  
125 The 40 clips in the survey were made up of two components. There was a set  
126 of 20 predetermined videos that were scored by everyone. The remaining 20 were  
127 determined by an adaptive threshold finding procedure, where the next clip shown  
128 depends on previous answers. Specifically, the truncated staircase method was  
129 used [13] to estimate thresholds (see below). To conceal this structure and reduce  
130 order effects these two components (predetermined set and threshold finding) have  
131 been mixed together such that each predetermined clip was followed by a clip used

---

<sup>1</sup>The survey can be reached at <http://groups.inf.ed.ac.uk/worms/index.html>

132 to detect the threshold. The participants were not told in advance of these two  
133 underlying components to eliminate possible cognitive biases.

134

135 To gather responses to the survey, we emailed 47 experts (PIs identified from  
136 publications on *C. elegans* behaviour) inviting them and their laboratory members  
137 to participate. The survey was also advertised through the social media presence  
138 of the OpenWorm project.

### 139 Selection of predetermined clips

140 To select the 20 predetermined clips, the eigenshape annotator (ESA) was used [3].  
141 In brief ESA is an unsupervised behavioural annotator that produces a probabilis-  
142 tic annotation. Events were selected that are labeled as Omega turns, but had  
143 a high entropy ( $0.75H_{max} \leq H$ ), i.e. Omega events were selected that had a high  
144 classification uncertainty. 158 events met this criteria and from this set 20 were  
145 selected randomly, see the online *Supplementary videos* to watch the clips.

### 146 Adaptive threshold finding

147 To deploy an adaptive threshold finding technique, it was necessary to have a single  
148 metric by which Omega turns could be ranked. We developed a ‘tightness’ metric  
149 score based on the Omega turn detection algorithms in the literature. Most Omega  
150 turn detection algorithms recognise such events when a certain feature exceeds a  
151 user defined threshold. Features that are commonly associated with Omega turns  
152 are solidity, midbody angle, head-tail distance and midbody bend. For a visual  
153 explanation for each of these features see Figure 1.

154

155 For each Omega event the peak amplitude of these features were measured.  
156 Across all events the z-score was calculated for each feature peak and the tight-  
157 ness score of each event is the mean z-score across the four features. This procedure  
158 ranks the Omega-like events from wide amplitude turns to the sharper, more ‘char-  
159 acteristic’ Omega turns. It is not claimed that the tightness score captures every  
160 variation of Omega like events. However it quantifies the sharpness of coils that  
161 is the key feature of turning behaviours. For a demonstration of the resulting  
162 ranking see the online *Supplemental Video 1*.

163

164 A truncated staircase method was used to estimate an expert’s omega detection  
165 threshold (measured on tightness score) [13]. The equation to select the next clip  
166 is

$$T_{n+1} = T_n - \delta(2R_n - 1) + z, \quad (3)$$

167 where  $\delta$  is a fixed step size (in tightness score),  $T_n$  is the tightness of the clip  
168 shown at the  $n^{\text{th}}$  step and  $R_n$  is the  $n^{\text{th}}$  response ( $R_n = 1$  if the answer is yes and  
169  $R_n = 0$  if the answer is no) and  $z$  is a small random variation to avoid repetitions.  
170 In this process the sequence of clips has either increasing or decreasing  $T_n$  until a  
171 switch in the subject's response (from yes to no, or no to yes) for successive clips  
172 occurs. In this case the step direction is reversed and again the stimulus strength  
173 ( $T_n$ ) monotonically increases or decreases until the next switch in response. To  
174 estimate the threshold, the average  $T_n$  at the points where the subject switched  
175 responses is taken.

## 176 3 Results

### 177 3.1 Consistency of Omega detection algorithms

178 The consistency of four Omega turn detection algorithms was quantified. In Table  
179 1 the precision, sensitivity and F-score of the methods are presented relative to  
180 each other. The scores are calculated first using the parameter settings originally  
181 provided, and then when the parameters of two methods were tuned for optimal  
182 match in outputs (results given in brackets; for details of the tuning procedure  
183 see the *Threshold tuning* section). Without tuning, the results show little consis-  
184 tency, with an average F-score of 0.3. Even with tuning to find the best match,  
185 the F-score frequently stays below 0.5, indicating poor consistency in classification.

186  
187 The Omega detection threshold was also estimated for each algorithm using  
188 the same methodology as for expert annotations (see *Adaptive threshold finding*).  
189 The results are shown on Figure 2B, for this figure the original parameters from  
190 the publications were used. Note that in agreement with Table 1 there is overlap  
191 in the confidence intervals, but there is no clear consensual threshold.

192  
193 The algorithm by [4] produces the worst match to the other algorithms. This  
194 is due to the method only picking out the sharpest of Omega turns, hence it iden-  
195 tifies many fewer events compared to the other methods. It is not argued that any  
196 of the methods assessed is worse or better than the others, but rather the point  
197 is that results could differ significantly depending on which method a particular  
198 analysis uses.

199

## 200 **3.2 Consistency of expert annotation**

201 Overall 27 survey responses were collected in the period 2016 May 30 - June 14.  
202 For the results presented here we have discarded the responses whose confidence  
203 in detecting Omega turns was below 4, so only expert annotation is analysed (19  
204 participants in total).

205  
206 As described in the *Methods*, the survey had two components: a set of prede-  
207 termined clips and an adaptive threshold finding procedure. Figure 2A shows the  
208 distribution of answers for the predetermined clips, which had been selected for  
209 high classification uncertainty according to an unsupervised behavioural annotator  
210 (see *Methods*). None received a unanimous consensus, and only 6 were judged the  
211 same by more than 75% (at least 15 out of 19) of the experts. Almost half the  
212 clips produced a split of 12:7 or worse.

213  
214 The estimated decision thresholds for each expert and the corresponding 95%  
215 confidence intervals are shown on Figure 2B. Note the different size of confidence  
216 intervals reflects the number of samples to estimate the threshold, which depends  
217 on the number of switch points from yes to no for each subject in the sequence of 20  
218 presentations (see *Adaptive threshold finding*). It is nevertheless also an indicator  
219 of the subject's (internal) consistency as more switch points, and hence smaller  
220 C.I., suggests the staircase quickly converged to oscillate around a specific value.  
221 It is clear that the estimated thresholds spread widely, with no region where the  
222 majority cluster, or all confidence intervals overlap.

## 223 **4 Discussion**

224 In this paper we have shown that both automated and expert annotations of  
225 *C. elegans*'s Omega turns are surprisingly divergent. First the implications for  
226 worm research are discussed. Then some general comments regarding supervised  
227 behavioural analysis is presented. Finally we speculate whether the observed an-  
228 notation inconsistency is a more general feature of behavioural studies.

229  
230 Characterising *C. elegans* behaviour often involves an estimate of Omega turn  
231 probability [3–7,9]. It is important to check whether the algorithms used to detect  
232 Omega turns are consistent, otherwise it is difficult to make cross-publication com-  
233 parisons. It was found that the four Omega turn detection algorithms we tested  
234 produce a surprisingly divergent annotation even after their respective parameters  
235 have been adjusted for optimal match.

236

237 One way to overcome the inconsistency problem would be if the community  
238 adopted the same platform for behavioural analysis. There is a range of publicly  
239 available packages [3–5], however, each comes with its own strengths and weak-  
240 nesses, hence it is difficult to see the whole community adopting any one of these  
241 methods. A potential solution would be an open-source software that is devel-  
242 oped and maintained not by a single laboratory, but rather by the whole research  
243 community. This way each lab would have ownership and the cross talk between  
244 laboratories could lead to a deeper appreciation of the limitations of each analysis  
245 technique.

246  
247 A potential source of the inconsistency we have observed is that the Omega turn  
248 is not a distinct behaviour, but rather a part of a spectrum of turning behaviours.  
249 We have previously argued for this possibility based on the high proportion of  
250 uncertain classification of behavioural events [3]. Others have also supported this  
251 hypothesis based on the geometry of locomotion states [14] and based on the con-  
252 tinuous neuronal representation of motor sequences [15].

253  
254 A major limitation of our work, in both our earlier paper and the current publi-  
255 cation, is that events could not be analysed where the worm was intersecting itself  
256 for an extended period (see *Methods*). Recently a method was developed that  
257 can resolve coiling postures [11]. Their analysis of eigenworm amplitudes found a  
258 multi-modal distribution that could be used as a data driven definition of Omega  
259 turns. Furthermore this study reports that ‘beyond’ Omega turns there is another  
260 sharper turning behaviour, the Delta turn.

261  
262 However one should note that in this study the experimental conditions were  
263 not identical to ours. In the CBD data (used here) worms are browsing in food,  
264 while in this study worms were analysed off food. The 1<sup>st</sup> and 3<sup>rd</sup> eigenworms  
265 switch position (sorted by eigenvalues) in these two conditions indicating that  
266 the behaviour is altered (off food the first two eigenworms are associated with  
267 locomotion and the 3<sup>rd</sup> one is associated with turns, on food the 1<sup>st</sup> eigenworm  
268 corresponds to the turning postures) [5,16]. Therefore the results may or may not  
269 generalise to other experimental conditions.

270  
271 Our analysis of expert annotation has general implications for supervised ap-  
272 proaches to behavioural analysis. The common element to these methods is that  
273 they take an investigator labeled dataset and then an algorithm learns to repro-  
274 duce the expert annotation [1]. As a consequence, supervised methods can be only  
275 as consistent as their training data. Therefore prior to using supervised methods  
276 we would urge investigators to first examine the variability of expert opinion. Fur-



277 furthermore we note that unsupervised methods are often evaluated against a human  
278 produced ‘ground truth’ dataset. This evaluation process imposes subjective fac-  
279 tors and hence leads to similar problems as with the supervised methods. The  
280 validation of unsupervised methods is a complex issue that raises many philosoph-  
281 ical questions as well [17, 18].

282

283 Although we have only analysed one specific behaviour of one model organ-  
284 ism, the observed inconsistencies in behavioural annotations (both expert and  
285 automated) seem likely to be more widespread. For example there is an analo-  
286 gous uncertainty about how to define the behavioural states of larval *Drosophila*  
287 *melanogaster* [3, 19–22]. Different publications use different ways of defining the  
288 behavioural states, most likely due to the difficulty in finding an unambiguous  
289 characterisation. As a result, a similar inconsistency of the various analysis tech-  
290 niques should be a cause for concern in reproducibility of maggot research. We  
291 hope that with our analysis we have inspired investigators to carefully look at the  
292 issue of consistency for other model organisms as well.

## 293 **Acknowledgements**

294 The authors would like to express their gratitude for Aidan Rocke for his initial  
295 work on this project. Furthermore we would like to thank Andre Brown, Emanuel  
296 Busch and members of the Insect Robotics group for their feedback on the survey  
297 prototype. This work was supported by grants EP/F500385/1 and BB/F529254/1  
298 for the University of Edinburgh School of Informatics Doctoral Training Centre  
299 in Neuroinformatics and Computational neuroscience from the UK Engineering  
300 and Physical Sciences Research Council (EPSRC), UK Biotechnology and Biolog-  
301 ical Sciences Research Council (BBSRC), and the UK Medical Research Council  
302 (MRC), and the FP7 program MINIMAL.

## 303 **Author contributions**

304 BS conceived the study, developed the code, analysed the data and wrote the arti-  
305 cle. TS developed the web implementation of the Omega event selection algorithm  
306 and maintained the survey’s website. BW has supervised the project and helped  
307 to write the manuscript.

308 **Figures**

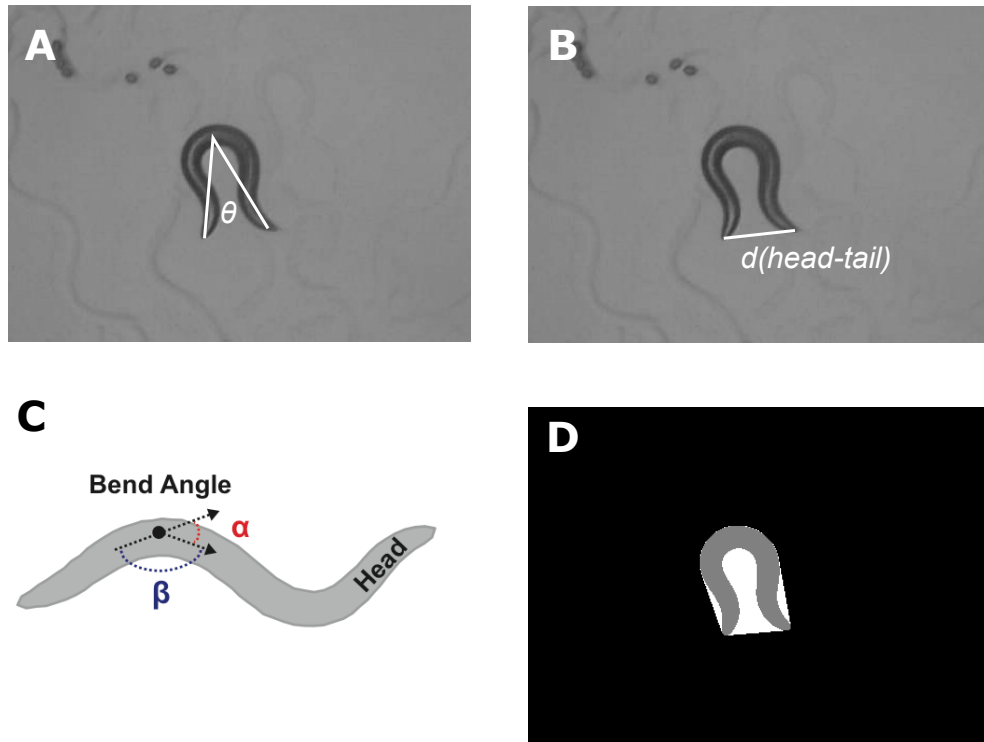


Figure 1: Visual explanation of the features that have been used to construct the tightness score. Panel **A** shows the midbody angle  $\theta$ , which is the angle between the head-middle and middle-tail vectors. Note that  $\pi - \theta$  is the angle of reorientation of the event [6]. Panel **B** shows the head-tail distance. **C** illustrates worm bending that is measured using the supplementary angles to the bends formed along the skeleton. The bend angle ( $\alpha$ ) is the difference in tangent angles at each point; or, alternatively phrased, the supplementary angle ( $\alpha$ ) with respect to the angle formed by any three consecutive points ( $\beta$ ). To detect Omega turns the midbody bend is calculated, which is the mean supplementary angle along the middle 1/3 of the worm's body (image and caption is taken from [5]). Finally panel **D** introduces solidity, a measure of the overall concavity. It is defined as the ratio of the image (the worm's body in grey) and the area of the convex hull (shown in white).

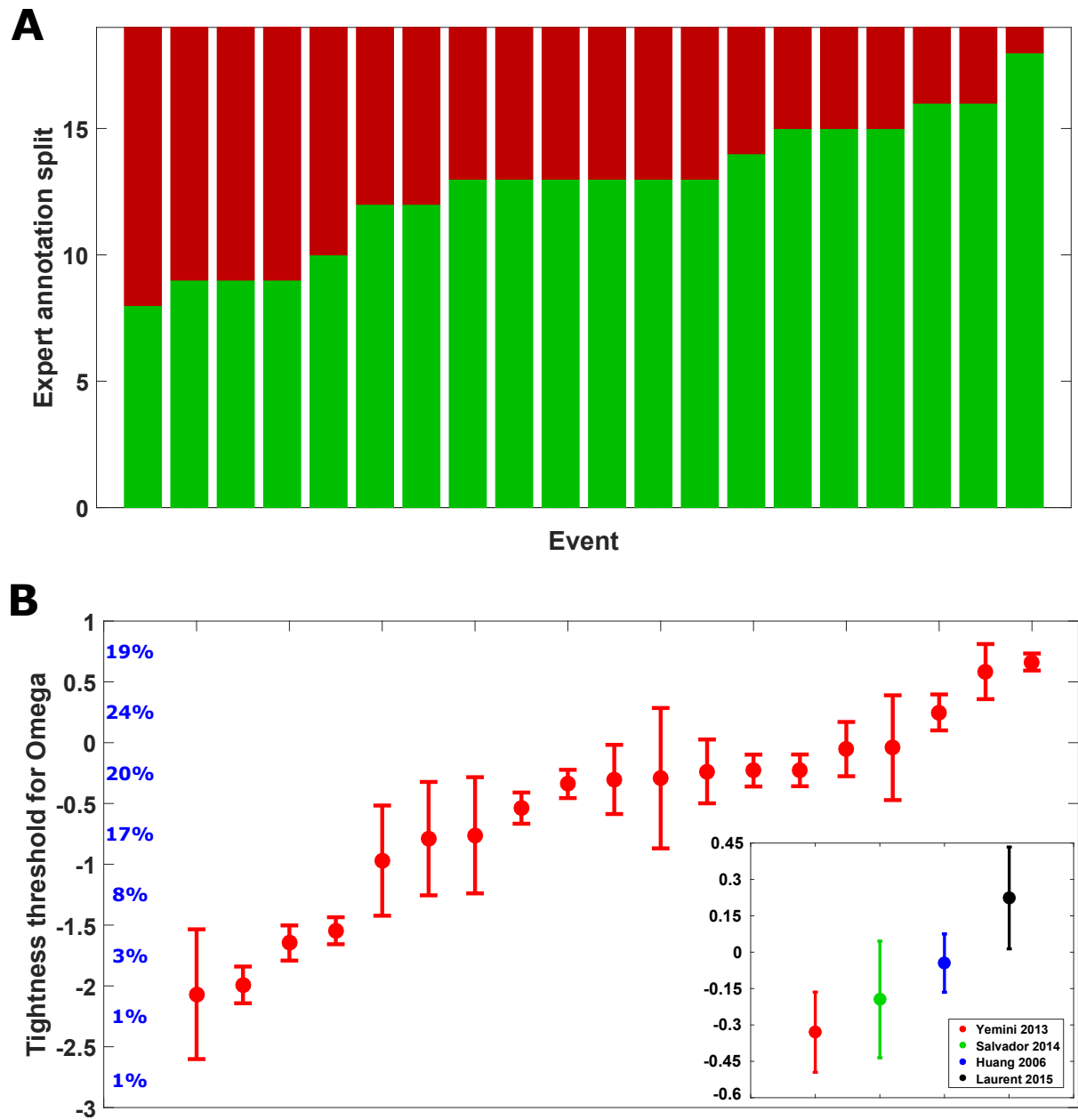


Figure 2: Outcomes of the Omega turn community survey. The data was filtered to exclude non-expert annotations, see the *Consistency of expert annotation* for details. Top panel shows the split of experts (green: ‘yes, it was an Omega’; red: ‘not an Omega’) for the set of 20 predetermined clips, ordered by the proportion of experts who agreed it was an Omega turn, which ranged from 8/19 to 18/19. Panel **B** shows the results of the threshold determination procedure. Each data point is one expert’s estimated tightness threshold to detect Omegas with the corresponding 95% confidence interval, ordered by increasing tightness. Inset shows the estimated threshold and confidence intervals for the Omega detection algorithms. Blue numbers next to the y-axis indicate what percentage of the data (all potential Omega events, see *Survey structure*) falls between tightness z-scores (e.g. 19% of the events had a tightness z-score between 0.5 and 1). This shows wide divergence in how many events different experts would classify as an Omega turn.

309 **Tables**

	Huang 2006	Yemini 2013	Salvador 2014	Laurent 2015
Huang 2006	1/1/1	0.40/0.46/0.43 (0.64/0.65/0.65)	0.28/0.15/0.20 (0.52/0.38/0.43)	0.13/0.67/0.22 (0.79/0.69/0.74)
Yemini 2013	0.46/0.4/0.42 (0.66/0.67/0.67)	1/1/1	0.45/0.22/0.29 (0.66/0.43/0.51)	0.05/0.21/0.08 (0.92/0.69/0.79)
Salvador 2014	0.15/0.28/0.20 (0.48/0.52/0.43)	0.26/0.5/0.34 (0.47/0.71/0.56)	1/1/1	0.12/0.1/0.11 (0.62/0.83/0.77)
Laurent 2015	0.68/0.13/0.22 (0.64/0.79/0.74)	0.22/0.05/0.1 (0.7/0.93/0.8)	0.62/0.1/0.13 (0.83/0.72/0.77)	1/1/1

Table 1: Consistency of Omega turn detection algorithms. The top of each column shows which algorithm was taken as reference and the rows correspond to the algorithm being compared to it. In each cell the *Precision/Sensitivity/F – score* are reported, for a description of these measures see the section *Consistency quantification*. The numbers in parentheses in each cell report the same statistics with thresholds tuned for optimal match, see the section *Threshold tuning* for further details.

## 310 References

- 311 [1] Mayank Kabra, Alice A Robie, Marta Rivera-Alba, Steven Branson, and  
312 Kristin Branson. JAABA: interactive machine learning for automatic an-  
313 notation of animal behavior. *Nature Methods*, 10(1):64–67, 2013.
- 314 [2] Alex Gomez-Marin, Joseph J Paton, Adam R Kampff, Rui M Costa, and  
315 Zachary F Mainen. Big behavioral data: psychology, ethology and the found-  
316 ations of neuroscience. *Nature Neuroscience*, 17(11):1455–1462, 2014.
- 317 [3] Balázs Szigeti, Ajinkya Deogade, and Barbara Webb. Searching for motifs in  
318 the behaviour of larval *Drosophila melanogaster* and *Caenorhabditis elegans*  
319 reveals continuity between behavioural states. *Journal of The Royal Society*  
320 *Interface*, 12(113):20150899, 2015.
- 321 [4] Patrick Laurent, Zoltan Soltesz, Geoffrey M Nelson, Changchun Chen, Fausto  
322 Arellano-Carbajal, Emmanuel Levy, and Mario de Bono. Decoding a neural  
323 circuit controlling global animal state in *C. elegans*. *eLife*, 4:e04241, 2015.
- 324 [5] Eviatar Yemini, Tadas Jucikas, Laura J Grundy, André EX Brown, and  
325 William R Schafer. A database of *Caenorhabditis elegans* behavioral phe-  
326 notypes. *Nature Methods*, 10(9):877–879, 2013.
- 327 [6] Kuang-Man Huang, Pamela Cosman, and William R Schafer. Machine vi-  
328 sion based detection of omega bends and reversals in *C. elegans*. *Journal of*  
329 *Neuroscience Methods*, 158(2):323–336, 2006.
- 330 [7] Liliana CM Salvador, Frederic Bartumeus, Simon A Levin, and William S  
331 Ryu. Mechanistic analysis of the search behaviour of *Caenorhabditis elegans*.  
332 *Journal of The Royal Society Interface*, 11(92):20131092, 2014.
- 333 [8] NEIL A CROLL. Components and patterns in the behaviour of the nematode  
334 *Caenorhabditis elegans*. *Journal of Zoology*, 176(2):159–176, 1975.
- 335 [9] Jonathan T Pierce-Shimomura, Thomas M Morse, and Shawn R Lockery. The  
336 fundamental role of pirouettes in *Caenorhabditis elegans* chemotaxis. *The*  
337 *Journal of Neuroscience*, 19(21):9557–9569, 1999.
- 338 [10] Dirk R Albrecht and Cornelia I Bargmann. High-content behavioral analysis  
339 of *Caenorhabditis elegans* in precise spatiotemporal chemical environments.  
340 *Nature Methods*, 8(7):599–605, 2011.
- 341 [11] Onno D Broekmans, Jarlath B Rodgers, William S Ryu, and Greg J Stephens.  
342 Resolving coiled shapes reveals new reorientation behaviors in *C. elegans*.  
343 *arXiv preprint arXiv:1603.04023*, 2016.

- 344 [12] David Martin Powers. Evaluation: from precision, recall and f-measure to  
345 roc, informedness, markedness and correlation. 2011.
- 346 [13] Bernhard Treutwein. Adaptive psychophysical procedures. *Vision Research*,  
347 35(17):2503–2522, 1995.
- 348 [14] Thomas Gallagher, Theresa Bjorness, Robert Greene, Young-Jai You, and  
349 Leon Avery. The geometry of locomotive behavioral states in *C. elegans*.  
350 *PloS One*, 8(3):e59865, 2013.
- 351 [15] Saul Kato, Harris S Kaplan, Tina Schrödel, Susanne Skora, Theodore H Lind-  
352 say, Eviatar Yemini, Shawn Lockery, and Manuel Zimmer. Global brain dy-  
353 namics embed the motor command sequence of *Caenorhabditis elegans*. *Cell*,  
354 163(3):656–669, 2015.
- 355 [16] Greg J Stephens, Bethany Johnson-Kerner, William Bialek, and William S  
356 Ryu. Dimensionality and dynamics in the behavior of *C. elegans*. *PLoS*  
357 *Computational Biology*, 4(4):e1000028, 2008.
- 358 [17] Jeremy G Todd, Jamey S Kain, and Benjamin de Bivort. Systematic explo-  
359 ration of unsupervised methods for mapping behavior. *bioRxiv*, page 051300,  
360 2016.
- 361 [18] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a  
362 review. *ACM Computing Surveys*, 31(3):264–323, 1999.
- 363 [19] Alex Gomez-Marin and Matthieu Louis. Multilevel control of run orientation  
364 in *Drosophila* larval chemotaxis. *Frontiers in Behavioral Neuroscience*, 8:38–  
365 38, 2013.
- 366 [20] Elizabeth A Kane, Marc Gershow, Bruno Afonso, Ivan Larderet, Mason  
367 Klein, Ashley R Carter, Benjamin L de Bivort, Simon G Sprecher, and Ar-  
368 avinthan DT Samuel. Sensorimotor structure of *Drosophila* larva phototaxis.  
369 *Proceedings of the National Academy of Sciences*, 110(40):E3868–E3877, 2013.
- 370 [21] Alex Gomez-Marin, Greg J Stephens, and Matthieu Louis. Active sampling  
371 and decision making in *Drosophila* chemotaxis. *Nature Communications*,  
372 2:441, 2011.
- 373 [22] CH Green, B Burnet, and KJ Connolly. Organization and patterns of inter-  
374 and intraspecific variation in the behaviour of *Drosophila* larvae. *Animal*  
375 *Behaviour*, 31(1):282–291, 1983.