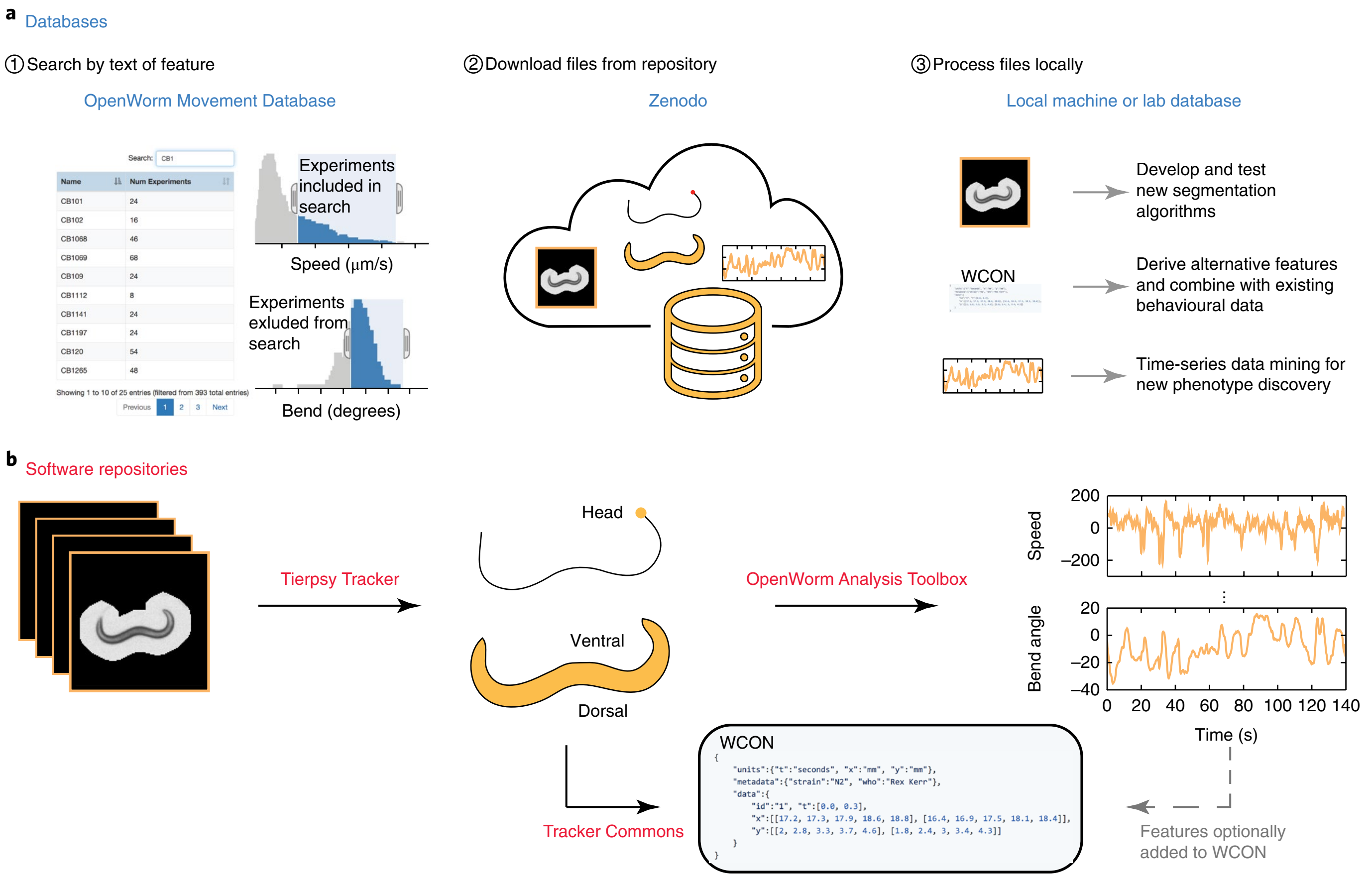
An open-source platform for analyzing and sharing worm-behavior data

To the Editor — Animal behavior is increasingly being recorded in systematic imaging studies that generate large datasets. To maximize the usefulness of these data, there is a need for improved resources for analyzing and sharing behavioral data that will encourage reanalysis and methodological developments[1]. However, for behavioral data, unlike genomic or protein structural data, there are no widely used standards. It is therefore desirable to make data available in a relatively raw form to enable flexibility in data analysis. For computational ethology to approach the level of maturity of other areas of bioinformatics, at least three challenges must be addressed: storing and accessing video files; defining flexible data formats to facilitate data sharing; and developing software to read, write, browse, and analyze the data. We have generated an open resource to begin addressing these challenges for *Caenorhabditis elegans* behavioral data.

To store video files and the associated features and metadata, we use a Zenodo.org community (an open-access repository for data) that provides durable storage and citability, and that supports contributions from other groups. We have also developed a web interface that enables filtering of the video files on the basis of feature histograms that can return, for example, fast and curved worms in addition to more standard searches for particular strains or genotypes (Fig. 1 and http://movement.openworm.org/). The database currently consists of 14,874 single-worm tracking experiments representing 386 genotypes (building on 9,203 experiments and 305 genotypes in a previous publication[2]) and includes data from several larval stages as well as data from aging experiments consisting of more than 2,700 videos of animals tracked daily from the L4 stage to death (Nature Research


a
Databases
① Search by text of feature
OpenWorm Movement Database
Experiments included in search
Speed (μm/s)
Experiments exluded from search
Bend (degrees)
② Download files from repository
Zenodo
③ Process files locally
Local machine or lab database
Develop and test new segmentation algorithms
WCON
Derive alternative features and combine with existing behavioural data
Time-series data mining for new phenotype discovery
b
Software repositories
Tierpsy Tracker
Head
Ventral
Dorsal
OpenWorm Analysis Toolbox
Speed
200
0
–200
Bend angle
20
0
–20
–40
0
20
40
60
80
100
120
140
Time (s)
Tracker Commons
WCON
Features optionally added to WCON

Search: CB1

Name	Num Experiments
CB101	24
CB102	16
CB1068	46
CB1069	68
CB109	24
CB1112	8
CB1141	24
CB1197	24
CB120	54
CB1265	48

Showing 1 to 10 of 25 entries (filtered from 393 total entries)

Previous 1 2 3 Next

```
{
    "units":{"t":"seconds", "x":"mm", "y":"mm"},
    "metadata":{"strain":"N2", "who":"Rex Kerr"},
    "data":{
        "id":"1", "t":[0.0, 0.3],
        "x":[[17.2, 17.3, 17.9, 18.6, 18.8], [16.4, 16.9, 17.5, 18.1, 18.4]],
        "y":[[2, 2.8, 3.3, 3.7, 4.6], [1.8, 2.4, 3, 3.4, 4.3]]
    }
}
```

Fig. 1 | Schematic of the searchable database and Tierpsy analysis pipeline. a, The OpenWorm Movement Database provides a web interface for searching by genotype, strain, and/or other discrete values, and interactive histograms with sliders to filter results on the basis of feature values. The interface points to data stored on Zenodo. The video and feature data can be further analyzed or combined with data collected by using other worm trackers through the Worm tracker Commons Object Notation (WCON), a human- and machine-readable JSON format. **b**, Tierpsy (short for tierpsychology, the German word for ethology) segments and tracks worms, extracting the outline and skeleton of each animal then determining the head–tail orientation. These data are saved in WCON. The OpenWorm Analysis Toolbox is then used to extract behavioral features.

Reporting Summary). Full-resolution videos are available in HDF5 containers that include gzip-compressed video frames, time stamps, worm outlines and midlines, feature data, and experimental metadata. HDF5 files are compatible with multiple languages including MATLAB, R, Python, and C. We have also developed an HDF5 video reader that allows video playback with adjustable speed and zoom (an important feature for reviewing high-resolution multiworm tracking data), as well as toggling of worm segmentation over the original video to verify segmentation accuracy during playback.

Second, we have defined an interchange format named Worm tracker Commons Object Notation (WCON), to facilitate data sharing and software reuse among groups working on worm behavior. WCON uses the widely supported JSON format to store tracking data as text that is readable by both humans and machines. It is compatible with single and multiworm[3] tracking data at any resolution, from a single point representing worm position over time[4] to many points representing the high-resolution skeleton of a moving worm[2]. It also supports custom feature additions so that individual laboratories can store their own specific datasets alongside the existing set of basic worm data. WCON readers are available for Python, MATLAB, Scala, and C. Detailed documentation for the file formats and software is available on the project page (https://github.com/openworm/tracker-commons/).

Finally, we have complemented the database and file formats with open-source software written in Python for single and multiworm tracking, feature extraction, review, and analysis (Supplementary Discussion; code and documentation in Supplementary Software or at https://doi.org/10.5281/zenodo.1323782, where compiled versions are also available).

The suite of tools reported here makes quantitative behavioral analysis and reanalysis accessible for both experimentalists and computational scientists. It may also serve as a template for similar efforts in other model-organism communities.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

Tierpsy Tracker is available as Supplementary Software and at https://doi.org/10.5281/zenodo.1323782. Updated versions will be made available at http://ver228.github.io/tierpsy-tracker/.

Data availability

Videos, skeleton (WCON) files, and feature files are available under a Creative Commons attribution (CC BY) license through the database page http://movement.openworm.org/ and Zenodo community page https://zenodo.org/communities/open-worm-movement-database/. ❐

Avelino Javer[1,2], Michael Currie[3], Chee Wai Lee[3], Jim Hokanson[3,4], Kezhi Li[1,2], Céline N. Martineau[5], Eviatar Yemini[6], Laura J. Grundy[7], Chris Li[8], QueeLim Ch'ng[9], William R. Schafer[7], Ellen A. A. Nollen[5], Rex Kerr[10] and André E. X. Brown[1,2]*

[1]MRC London Institute of Medical Sciences, London, UK. [2]Institute of Clinical Sciences, Imperial College London, London, UK. [3]OpenWorm Foundation, San Diego, CA, USA. [4]Department of Biomedical Engineering, Duke University, Durham, NC, USA. [5]European Research Institute for the Biology of Ageing, University of Groningen, Groningen, the Netherlands. [6]Department of Biological Sciences, Columbia University, New York, NY, USA. [7]MRC Laboratory of Molecular Biology, Cambridge, UK. [8]Department of Biology, City College of the City University of New York, New York, NY, USA. [9]Centre for Developmental Neurobiology, King's College London, London, UK. [10]Calico Life Sciences LLC, South San Francisco, CA, USA.
**e-mail: andre.brown@imperial.ac.uk*

Published online: 31 August 2018
https://doi.org/10.1038/s41592-018-0112-1

References

1. Gomez-Marin, A., Paton, J. J., Kampff, A. R., Costa, R. M. & Mainen, Z. F. *Nat. Neurosci.* **17**, 1455–1462 (2014).
2. Yemini, E., Jucikas, T., Grundy, L. J., Brown, A. E. X. & Schafer, W. R. *Nat. Methods* **10**, 877–879 (2013).
3. Swierczek, N. A., Giles, A. C., Rankin, C. H. & Kerr, R. A. *Nat. Methods* **8**, 592–598 (2011).
4. Ramot, D., Johnson, B. E., Berry, T. L. Jr., Carnell, L. & Goodman, M. B. *PLoS One* **3**, e2208 (2008).

Acknowledgements

This work was supported by the MRC through grant MC-A658-5TY30 to A.E.X.B. Q.C. is supported by an ERC Starting Grant (NeuroAge 242666), a Research Councils UK Fellowship, and the University of London Central Research Fund. Some strains were provided by the CGC, which is funded by the NIH Office of Research Infrastructure Programs (P40 OD010440).

Author contributions

A.J. wrote Tierpsy Tracker and analyzed data; M.C. wrote WCON viewer, the database, and OpenWorm Analysis Toolbox; C.W.L. wrote the database, web interface, and WCON viewer; J.H. wrote the MATLAB WCON viewer and OpenWorm Analysis Toolbox; K.L. wrote stage-alignment code; C.N.M. collected data; E.Y. wrote the skeletonization algorithm and stage-alignment code; L.J.G. collected data; C.L. contributed strains and planned experiments; Q.C. contributed strains and planned experiments; W.R.S. planned the study; E.A.A.N. contributed strains and planned experiments; R.K. designed WCON and wrote several readers; A.E.X.B. planned the study and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0112-1.

nature research

Corresponding author(s): André Brown

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The single worm tracking data was collected using WormTracker 2.0 (custom code written in Java) as described previously in Yemini et al. (2013) Nature Methods. The multiworm data (Fig. S5) was collected using Gecko version 2.0.3.1 to capture data from the cameras (http://gecko.visionexperts.co.uk/).
Data analysis	All tracking and feature extraction was performed using custom written code available at http://ver228.github.io/tierpsy-tracker/. The classification results reported in Fig. S5 were computed using the open source library PyTorch.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Videos, skeleton (WCON) files, and feature files are available with a Creative Commons attribution (CC BY) license through the database page http://movement.openworm.org/ and Zenodo community page https://zenodo.org/communities/open-worm-movement-database/

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences　　□ Behavioural & social sciences　　□ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were chosen to be approximately 20 worms per strain. This provides a power of 0.8 to detect a ~1 standard deviation effect (see Yemini et al. (2013) Fig. S3).
Data exclusions	Data were excluded if worms were lost during tracking or if fewer than 100 frames were skeletonized (this corresponds to less than 0.05% of frames being skeletonised) to exclude severely under-sampled worms. Data were also excluded if a filename contained an error (for example, a non-existent gene name) that could not be reliably corrected with reference to lab notebooks since these data cannot be associated with a strain and therefore compared to other strains. Data were also excluded if the pipeline failed to complete for a given video (e.g. due to a corrupted video) since no feature data are available to analyze in this case. These exclusions were not pre-established before analysis.
Replication	No replication was performed.
Randomization	The tracker used for collecting control data from the reference strain (N2) was varied from day-to-day.
Blinding	No blinding was performed because the same features were extracted and the same analysis performed automatically regardless of strain identify. There was thus low risk of experimenter bias affecting the results.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
□	☒	Animals and other organisms
☒	□	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	This study involved both hermaphrodites and a small number of male C. elegans. Most data were collected from day 0 adults, but other ages are included and noted in the database. Over 300 strains were used. Strain identify is recorded in the database and presented with analysis results in the paper.

Wild animals	Wild isolates (Fig. S5) were obtained from the C. elegans Natural Diversity Resource (CeNDR). Details on the strains are available from the CeNDR page: https://www.elegansvariation.org/
Field-collected samples	No field samples were used in this study.