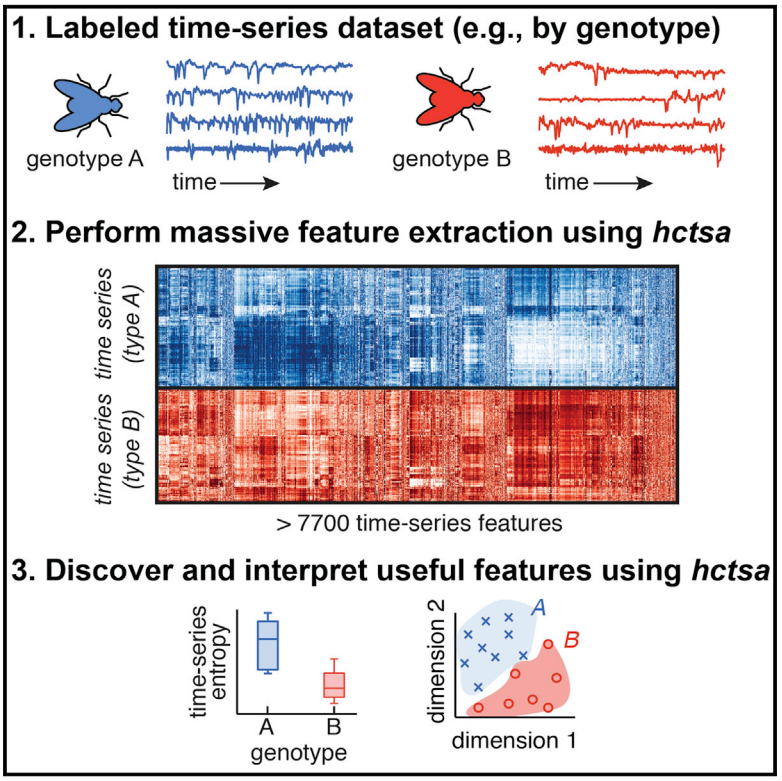
Tool

Cell Systems

hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction

Graphical Abstract



Highlights

- Fully documented and comprehensively tested software framework, *hctsa*
- Automatically identify interpretable quantitative phenotypes from time-series data
- Uses over 7,700 features from scientific time-series analysis literature
- Provides biological understanding from *C. elegans* and *Drosophila* movement data

Authors

Ben D. Fulcher, Nick S. Jones

Correspondence

ben.d.fulcher@gmail.com (B.D.F.),
nick.jones@imperial.ac.uk (N.S.J.)

In Brief

A new software tool, *hctsa*, uses massive feature extraction to automatically identify informative and interpretable quantitative phenotypes from time-series data.

Fulcher & Jones, 2017, Cell Systems *5*, 527–531
November 22, 2017 © 2017 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2017.10.001

CellPress

Cell Systems
Tool

OPEN ACCESS
CellPress

hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction

Ben D. Fulcher[1,2,4,5,*] and Nick S. Jones[3,*]
[1]Monash Institute of Cognitive and Clinical Neurosciences (MICCN), Monash University, Wellington Road, Clayton, VIC, 3800, Australia
[2]School of Physics, Sydney University, Physics Road, Camperdown, NSW, 2006, Australia
[3]Mathematics Department, Imperial College London, Huxley Building, Queen's Gate, London SW7 2AZ, UK
[4]Twitter: @bendfulcher
[5]Lead Contact
*Correspondence: ben.d.fulcher@gmail.com (B.D.F.), nick.jones@imperial.ac.uk (N.S.J.)
https://doi.org/10.1016/j.cels.2017.10.001

SUMMARY

Phenotype measurements frequently take the form of time series, but we currently lack a systematic method for relating these complex data streams to scientifically meaningful outcomes, such as relating the movement dynamics of organisms to their genotype or measurements of brain dynamics of a patient to their disease diagnosis. Previous work addressed this problem by comparing implementations of thousands of diverse scientific time-series analysis methods in an approach termed highly comparative time-series analysis. Here, we introduce *hctsa*, a software tool for applying this methodological approach to data. *hctsa* includes an architecture for computing over 7,700 time-series features and a suite of analysis and visualization algorithms to automatically select useful and interpretable time-series features for a given application. Using exemplar applications to high-throughput phenotyping experiments, we show how *hctsa* allows researchers to leverage decades of time-series research to quantify and understand informative structure in time-series data.

Time series, repeated measurements of a quantity taken through time, are recorded in increasing volumes in biology and medicine. This wealth of data has opened the door to a range of new research problems, including diagnosis of pathology from biomedical data streams in human patients (Hripcsak and Albers, 2013; Insel et al., 2010), understanding the role of specific neural circuits for behavior (Vogelstein et al., 2014), and linking genotype to phenotype to understand gene function (Nolan et al., 2000; Brown et al., 2013; Kain et al., 2013) or disease processes (Johnson et al., 2006; Gates et al., 2011; Yang et al., 2014). While differences in scalar phenotypes are relatively simple to calculate (such as the body length of a worm or the blood pressure of a human subject), it is less clear how to compare complex time-varying data streams (such as the movement dynamics of a worm, the heart rate fluctuations of a clinical patient, or the sequence of reaction times across a cognitive task). These diverse applications require a method for reducing complex time-series data streams to informative, low-dimensional summaries.

A common way of summarizing a time series is by measuring a simple statistic such as its sample mean, which has the advantage of being easily interpretable; e.g., knocking out the gene *unc-9* decreases the mean movement speed of the nematode worm, *Caenorhabditis elegans* (Yemini et al., 2013). However, this approach fails for many real-world applications in which the phenotypic differences are more subtle than simple mean shifts. Sophisticated tools for measuring structure in time-series data have been developed by a broad range of researchers, including contributions from the fields of statistics, electrical engineering, economics, statistical physics, dynamical systems, and biomedicine. This interdisciplinary literature includes summaries of the distribution of values in the data (e.g., Gaussianity, properties of outliers), autocorrelation structure (including power spectral measures), stationarity (how properties change over time), information theoretic measures of entropy and temporal predictability, linear and nonlinear model fits to the data, and methods from the physical nonlinear time-series analysis literature (Fulcher et al., 2013). There is currently no systematic way of leveraging this giant corpus of scientific work to determine which of these thousands of possible summary statistics best address a particular scientific hypothesis because the methods have typically been locked in discipline-specific journal articles.

Our previous work introduced the approach of highly comparative time-series analysis, in which the interdisciplinary time-series analysis literature is represented algorithmically in the form of thousands of features, each of which captures a different type of interpretable structure in a univariate time series. Comparing the performance of these features on a given dataset facilitates data-driven, statistically controlled selection of informative time-series summary statistics for phenotyping applications, overcoming an otherwise time-consuming and subjective manual task (Fulcher et al., 2013; Fulcher and Jones, 2014). A preliminary set of time-series feature extraction functions were made available with previous work (www.comp-engine.org/timeseries), but an accompanying platform for leveraging these features to tackle time-series analysis problems was missing. Here, we describe a refined version of our original

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

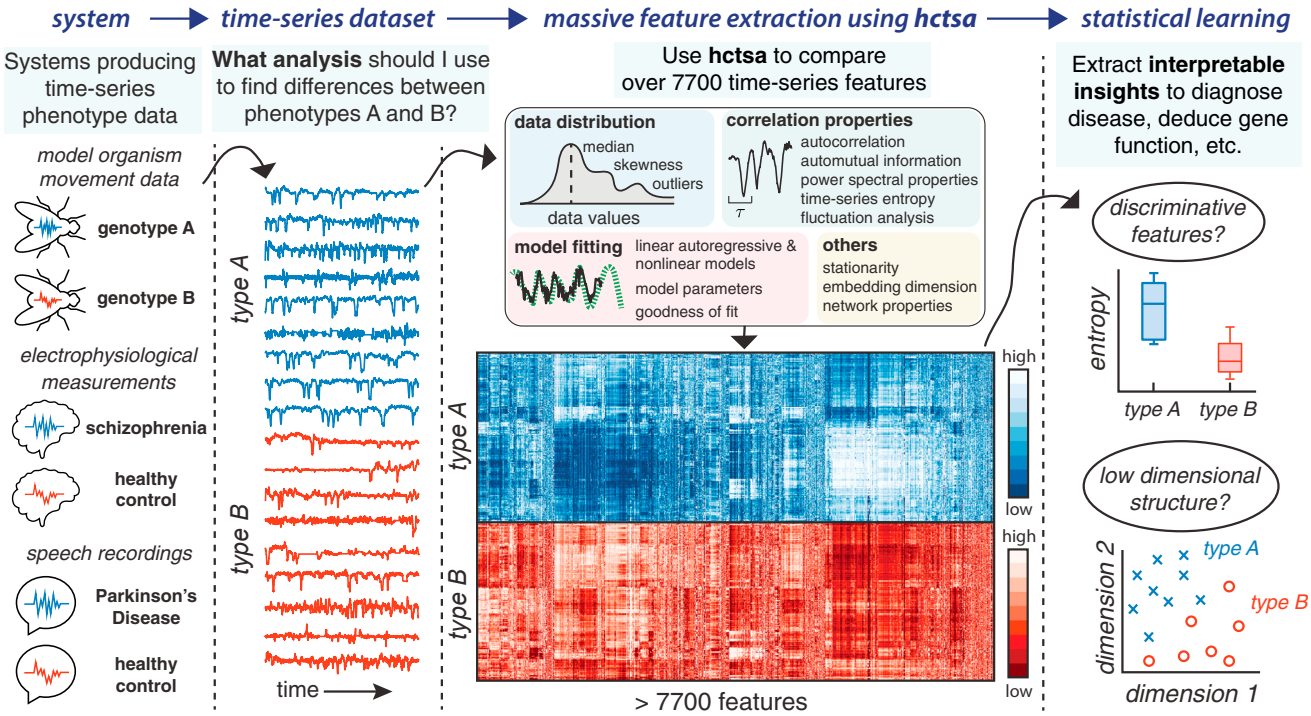


Figure 1. Using a Massive Interdisciplinary Library of Time-Series Analysis Methods to Quantify and Interpret Phenotypic Difference Using *hctsa*

We illustrate the problem of distinguishing two labeled classes of systems using measured time-series data. The *hctsa* package facilitates massive feature extraction to compare over 7,700 features of each time series, derived from an interdisciplinary time-series analysis literature. The feature matrix contains the result of this feature extraction, where each row represents a time series and each column represents a feature that encapsulates some property of that time series (e.g., measures of its autocorrelation structure, entropy, etc.). Colored (blue and red) labels the two types of data—e.g., electrophysiological recordings from healthy controls (A) or people with schizophrenia (B)—and dark/light labels low/high values of each feature, revealing rich structure in the dynamical properties of the dataset. A range of analysis functions are also included with *hctsa*, including those for learning interpretable differences between the labeled groups (visualized as a boxplot revealing that time series of type A have increased entropy), and visualizing informative low-dimensional structure in the dataset.

feature set (Supplemental Information) and introduce *hctsa*, a MATLAB-based software implementation of our methodology.

Given a time-series dataset, *hctsa* allows researchers to perform massive feature extraction, transforming each time series to a set of over 7,700 features that each encode a different scientific analysis method. The result can be visualized in *hctsa* as a feature matrix with a row for every time series and a column for every feature (Figure 1). The rich structure of this feature matrix reveals some features (i.e., areas of the scientific time-series analysis literature) that capture meaningful differences between different types of time series (e.g., lighter color for type A and darker color for type B in Figure 1) and thus represent promising candidates as quantitative phenotypes for distinguishing data of the two types. For demonstration purposes, we focus on distinguishing time series recorded from two different classes (e.g., a patient group and a control group), although the same general framework applies to multiclass classification or regression problems (Fulcher et al., 2013).

hctsa also includes a comprehensive suite of analytics for understanding structure in a time-series dataset, including: (1) identifying scientific methods that best quantify differences between labeled groups of data, providing interpretable insights into the phenotypic differences (incorporating permutation testing to statistically control for multiple hypothesis testing), (2) building classifiers that draw on the full diversity of time-series features to optimize the accuracy of phenotypic classification, and (3) visualizing low-dimensional structure in the dataset to understand potential clustering structure or other relationships between the time series. The *hctsa* package thus allows researchers to apply highly comparative time-series analysis to their own datasets, leveraging a comprehensive interdisciplinary literature on time-series analysis to gain interpretable and useful understanding of their data.

To demonstrate the software, we applied *hctsa* to movement speed time series of five different strains of *C. elegans* (Brown et al., 2013). Being short, noisy empirical recordings with no clear visual differences between strains, it is unclear what types of analysis methods might capture differences between the genotypes (Figure 2A, top). We used *hctsa* to compute >7,700 time-series features (subsequently filtered down to 6,504 well-behaved features, see STAR Methods), and used these features to predict genotype, obtaining a 10-fold cross-validated balanced accuracy of 80% (using a linear support vector machine [SVM]; chance level, 20%). We next used *hctsa* to identify 4,499 features of movement speed data that are individually informative of genotype ($q < 0.05$, using permutation testing with false-discovery-rate correction to control for multiple hypothesis testing). The 40 most informative features

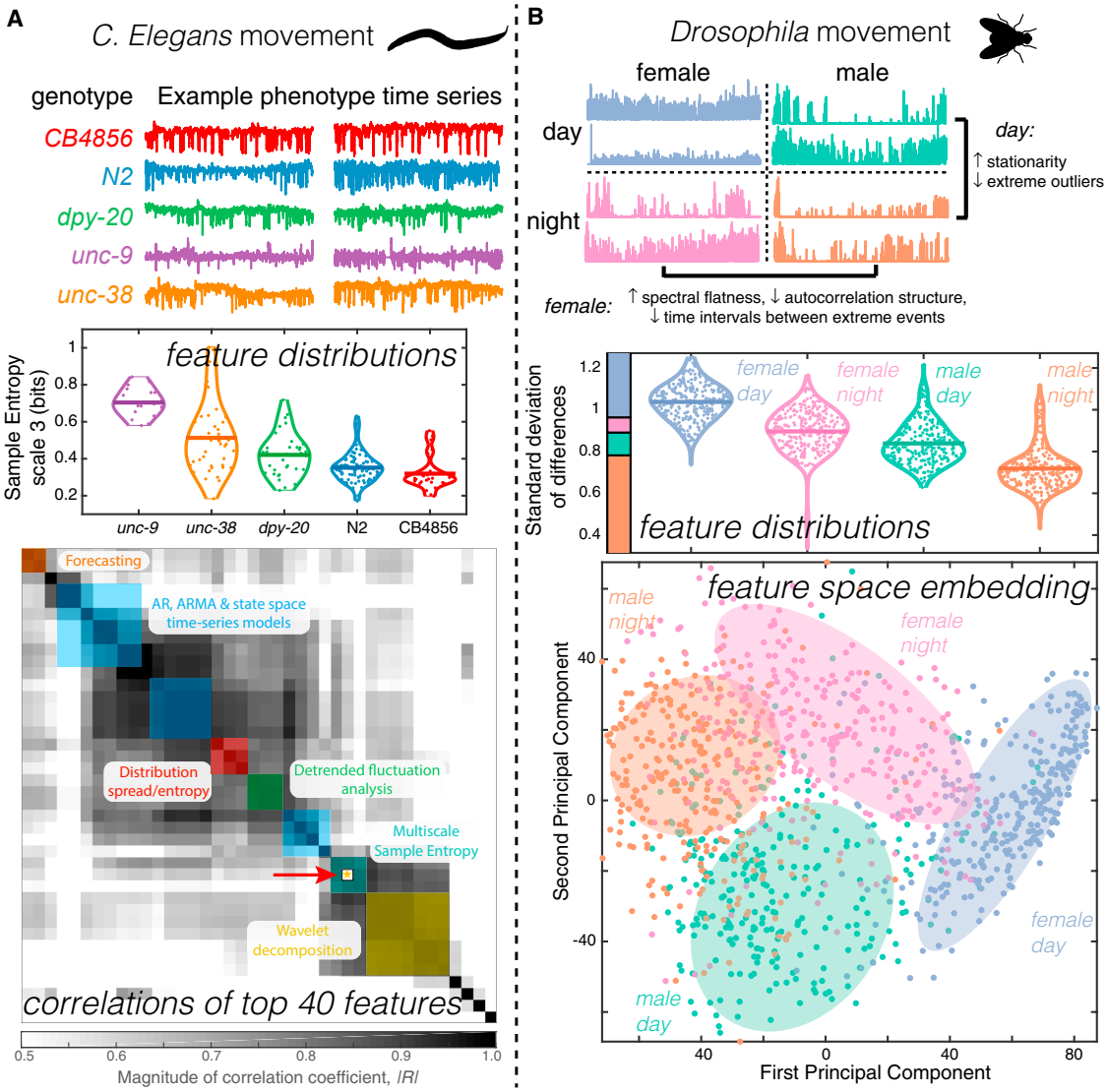


Figure 2. *hctsa* Uncovers Interpretable, Quantitative Phenotypic Differences in Movement-Speed Time Series of *C. elegans* and *Drosophila*

(A) *C. elegans*. Upper: Two examples of movement speed time series are shown for each of five genotypes. Middle: Class distributions of multiscale Sample Entropy, selected by *hctsa* as an informative measure, are shown as a violin plot, demonstrating that the neural mutant *unc-9* genotype has the highest average Sample Entropy at this scale, followed by *unc-38*, the morphological mutant *dpy-20*, the lab-based strain N2, and the wild-type strain CB4856. Lower: The top 40 features identified by *hctsa* for distinguishing the five genotypes span a wide range of time-series analysis techniques, labeled using color, and form sets of highly correlated groups. The multiscale Sample Entropy shown above is indicated with a red arrow and star.

(B) *Drosophila*. Upper: Two examples of movement speed time series are shown for each of four groups, labeled as either male or female, and either day or night. Interpretable measures of difference between each pair of conditions were extracted using *hctsa* and are summarized using text. Middle: *hctsa* identified the SD of successive changes in movement speed as a simple but highly discriminative feature, shown as a violin plot. Lower: A two-dimensional principal components projection of the dataset across the full *hctsa* feature library is informative of the class structure in the dataset. Shading has been added to guide the eye.

span diverse methodological literature, including autoregressive and state space model-fitting methods, detrended fluctuation analysis, local mean forecasting, multiscale Sample Entropy, and wavelet decompositions of the signal. A plot of the structured pairwise correlation between features allows the user to visually identify how different types of methods capture qualitatively different types of important time-series structure (Figure 2A, lower).

hctsa provides tools for investigating informative individual features in more detail. For example, a violin plot of Sample Entropy, *SampEn*(2,0.15), computed across 100 ms bins reveals physiologically interpretable differences between the genotypes (Figure 2A, middle). Sample Entropy quantifies the "unpredictability" of the time series at a given timescale (Costa et al., 2005). The two neural mutants *unc-38* (which encodes a nicotinic acetylcholine receptor alpha subunit) and *unc-9* (which encodes a structural component of gap junctions) have similar mean movement speeds (data not shown), but our analysis suggests that they affect movement distinctly, exhibiting significant differences in their movement predictability. The selection of

this multiscale entropy measure as a quantitative phenotype for *C. elegans* movement mirrors detailed manual research proposing the similar concept of "compressibility" of posture sequences as a quantitative phenotype for *C. elegans* (Gomez-Marin et al., 2016).

To demonstrate the flexibility of *hctsa* in extracting quantitative phenotypes, we also applied it to 12 hr *Drosophila* movement speed time series, labeled as either day (light on) or night (light off), and as either male or female (Figure 2B) (Gilestro, 2012; Geissmann et al., 2017). *hctsa* successfully distinguishes day versus night recordings (with a mean 10-fold cross-validated balanced accuracy of 98%), predicts the sex of the organism (96%), and classifies the four combination classes (colored in Figure 2B; 95%). *hctsa* selects different time-series features for different groupings of the data (labeled in Figure 2B, upper), capturing the less predictable movement in females than males (increased spectral flatness), and more bursty sleep/activity dynamics at night (increased temporal stationarity). The spread of incremental differences in the *Z*-scored time series is a simple measure of temporal predictability found by *hctsa* that is increased during the day and in females (Figure 2B, middle). *hctsa* thus goes beyond simple comparisons of the overall amount of movement—e.g., females have shorter sleep bouts than males (Gilestro, 2012)—by identifying more subtle measures of sexually dimorphic behavior, including movement predictability (reduced in females) and time intervals between large movements (reduced in females). More erratic female *Drosophila* movement may reflect their need to forage for food and select egg-laying sites, in contrast to the more predictable male behavior of conserving energy to avoid predators (Isaac et al., 2010). *hctsa* can also be used to structure a time-series dataset in a low-dimensional representation of the combined behavior of thousands of time-series features. This unsupervised analysis clearly separates different types of *Drosophila* movement (Figure 2B, lower).

In summary, *hctsa* automates the selection of quantitative phenotypes from time-series data by leveraging a large and interdisciplinary time-series analysis literature. The software allows researchers to distill a large methodological literature down to those methods that are most informative for the problem at hand, directing them to subsequently understand and properly interpret these methods in the context of their domain application. In addition to the two phenotyping case studies demonstrated here, *hctsa* has general utility, including behavioral phenotyping in cognitive science and diagnosis of disease from biomedical data streams such as heart rates or brain dynamics. Furthermore, although we focus on classification problems here, we note that the same approach applies to regression problems, where one aims to find time-series features that vary with a continuous variable (such as the dosage of a drug, a standardized depression score of a patient, etc.) (Fulcher et al., 2013). *hctsa* is available at www.github.com/benfulcher/hctsa, with accompanying comprehensive documentation at www.gitbook.com/book/benfulcher/hctsa-manual.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information includes the list of code files used in *hctsa* and can be found with this article online at https://doi.org/10.1016/j.cels.2017.10.001.

AUTHOR CONTRIBUTIONS

Conceptualization, B.D.F. and N.S.J.; Methodology, B.D.F. and N.S.J.; Software, B.D.F.; Formal Analysis, B.D.F. and N.S.J.; Data Curation, B.D.F.; Writing – Original Draft, B.D.F.; Writing – Review & Editing, B.D.F. and N.S.J.; Visualization, B.D.F.; Supervision, B.D.F. and N.S.J.; Resources, N.S.J. and B.D.F.

ACKNOWLEDGMENTS

We thank Andre Brown and Bertalan Gyenes for sharing the *C. elegans* movement dataset and helpful feedback on the resulting analysis and manuscript. We thank Giorgio Gilestro and Quentin Geissmann for sharing the *Drosophila* movement dataset and helpful feedback on the resulting analysis. Many thanks to Rachael Fulcher for help with graphic design, and to Alex Fornito and Iain Johnston for useful feedback on the manuscript. Some *C. elegans* strains were provided by the *Caenorhabditis* Genetics Center (CGC), which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). B.D.F. was supported by an NHMRC Early Career Fellowship (no. 1089718). N.S.J. was supported by an EPSRC grant (EP/N014529/1). B.D.F. and N.S.J. are founders of Engine Analytics Pty Ltd and members of its scientific advisory board.

Received: November 24, 2016
Revised: May 23, 2017
Accepted: September 28, 2017
Published: November 1, 2017

REFERENCES

Brown, A.E.X., Yemini, E.I., Grundy, L.J., Jucikas, T., and Schafer, W.R. (2013). A dictionary of behavioral motifs reveals clusters of genes affecting *Caenorhabditis elegans* locomotion. Proc. Natl. Acad. Sci. USA *110*, 791–796.

Costa, M., Goldberger, A.L., and Peng, C.K. (2005). Multiscale entropy analysis of biological signals. Phys. Rev. E *71*, 021906.

Donelson, N., Kim, E.Z., Slawson, J.B., Vecsey, C.G., Huber, R., and Griffith, L.C. (2012). High-resolution positional tracking for long-term analysis of *Drosophila* sleep and locomotion using the "Tracker" program. PLoS One *7*, e37250.

Fulcher, B.D., Georgieva, A.E., Redman, C.W.G., and Jones, N.S. (2012). Highly comparative fetal heart rate analysis. Conf. Proc. IEEE Eng. Med. Biol. Soc. *2012*, 3135–3138.

Fulcher, B.D., Little, M.A., and Jones, N.S. (2013). Highly comparative time-series analysis: the empirical structure of time series and their methods. J. Roy. Soc. Interf. *10*, 20130048.

Fulcher, B.D., and Jones, N.S. (2014). Highly comparative feature-based time-series classification. IEEE Trans. Knowl. Data Eng. *26*, 3026–3037.

Gates, H., Mallon, A.-M., and Brown, S.D.M. (2011). High-throughput mouse phenotyping. Methods *53*, 394–404.

Geissmann, Q., Rodriguez, L.G., Beckwith, E.J., French, A.S., Jamasb, A.R., and Gilestro, G.F. (2017). Ethoscopes: an open platform for high-throughput ethomics. PLoS Biol. *15*, e2003026.

Gilestro, G.F. (2012). Video tracking and analysis of sleep in *Drosophila melanogaster*. Nat. Protoc. *7*, 995.

Gomez-Marin, A., Stephens, G.J., and Brown, A.E.X. (2016). Hierarchical compression of *C. elegans* locomotion reveals phenotypic differences in the organisation of behaviour. bioRxiv. https://doi.org/10.1101/029462.

Hripcsak, G., and Albers, D.J. (2013). Next-generation phenotyping of electronic health records. J. Am. Med. Inform. Assoc. *20*, 117–121.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research Domain Criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry *167*, 748–751.

Isaac, R.E., Li, C., Leedale, A.E., and Shirras, A.D. (2010). *Drosophila* male sex peptide inhibits siesta sleep and promotes locomotor activity in the post-mated female. Proc. R. Soc. Lond. B *277*, 65–70.

Johnson, J.T., Hansen, M.S., Wu, I., Healy, L.J., Johnson, C.R., Jones, G.M., Capecchi, M.R., and Keller, C. (2006). Virtual histology of transgenic mouse embryos for high-throughput phenotyping. PLos Genet. *2*, e61.

Kain, J., Stokes, C., Gaudry, Q., Song, X., Foley, J., Wilson, R., and de Bivort, B. (2013). Leg-tracking and automated behavioural classification in *Drosophila*. Nat. Comm. *4*, 1910.

Nolan, P.M., Peters, J., Strivens, M., Rogers, D., Hagan, J., Spurr, N., Gray, I.C., Vizor, L., Brooker, D., Whitehill, E., et al. (2000). A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. Nat. Genet. *25*, 440–443.

Vogelstein, J.T., Park, Y., Ohyama, T., Kerr, R.A., Truman, J.W., Priebe, C.E., and Zlatic, M. (2014). Discovery of brainwide neural-behavioral maps via multi-scale unsupervised structure learning. Science *344*, 386–392.

Yang, B., Treweek, J.B., Kulkarni, R.P., Deverman, B.E., Chen, C.-K., Lubeck, E., Shah, S., Cai, L., and Gradinaru, V. (2014). Single-cell phenotyping within transparent intact tissue through whole-body clearing. Cell *158*, 945–958.

Yemini, E., Jucikas, T., Grundy, L.J., Brown, A.E.X., and Schafer, W.R. (2013). A database of *Caenorhabditis elegans* behavioral phenotypes. Nat. Methods *10*, 877–879.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Time-series data and *hctsa* output files for *C. elegans*	This paper	Figshare: https://dx.doi.org/10.4225/03/580478f951263
Time-series data and *hctsa* output files for *Drosophila*	This paper	Figshare: https://dx.doi.org/10.4225/03/5804798d2a2ec
Experimental Models: Organisms/Strains		
C. elegans: Wild-type strains: CB4856 and N2. Mutant strains are *dpy-20(e1282)*, *unc-9(e101)*, *unc-38(e264)*	Mutant strains from *Caenorhabditis* Genetics Center using EMS in genomic mutagenesis screens	WormBase IDs: CB4856, N2, CB1282, CB101, CB904
D. melanogaster: control line *CantonS*	Bloomington *Drosophila* Stock Center	http://flystocks.bio.indiana.edu/
Software and Algorithms		
MATLAB 2016b	The MathWorks, Natick, MA	https://mathworks.com/products/matlab.html
hctsa	This paper	https://github.com/benfulcher/hctsa

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Ben Fulcher (ben.d.fulcher@gmail.com).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

In this work we apply *hctsa* to two different datasets, which are described below.

Caenorhabditis elegans Movement Speed Data

Movement speed time-series data were obtained from approximately 15 min videos obtained using tracking microscopes (Brown et al., 2013; Yemini et al., 2013) (see wormbase.org for more information). A total of 226 movement time series sampled at 30.03 Hz were obtained from the CB4856 (Hawaiian wild isolate, 29 time series) and N2 (lab strain, 100 time series) strains, and the mutants *dpy-20(e1282)* (34 time series), *unc-9(e101)* (20 time series), *unc-38(e264)* (43 time series). For the *dpy-20*, *unc-9*, *unc-38* knockouts and the CB4856 strain, all available data at the specified frame rate were used. For the wildtype N2 strain, we took a random sample of 100 of the 1 200 time series recorded at a sampling rate of 30.03 Hz. If missing data in a time series made up less than 15% of its length and in a contiguous block at the beginning or end of the recording, the time series was retained with this section of missing data removed, otherwise the time series was removed.

Drosophila melanogaster Movement Speed Data

We analyzed time series of the movement speed of flies restricted to a one-dimensional tube and tracked continuously for between 3 and 6 days using video tracking (Gilestro, 2012; Donelson et al., 2012; Geissmann et al., 2017). Movement speed was estimated as the maximum speed of the measured data (sampled at approximately 2 Hz) in each non-overlapping 10 s time window, where displacements are measured as the Euclidean distance between successive co-ordinates of the fly. In this way, here we analyze these time series of movement speed, sampled at a rate of 0.1 Hz. Time series were split into 12 hr segments and labeled as either 'day' (light on, 574 time series) or 'night' (light off, 574 time series), and as either 'male' (554 time series) or 'female' (594 time series).

METHOD DETAILS

Following from the original concept and proof of principle for a highly comparative approach to time-series analysis (Fulcher et al., 2013), this article introduces a well-documented and user-friendly Matlab-based software platform for performing it (MATLAB is a product of The MathWorks, Natick, MA). The set of over 7,700 features has been developed and refined through applications to a wide range of research and industrial problems over many years. A full analysis pipeline has also been built to allow researchers to run highly comparative analysis on their own data, including functions for initiating new analysis tasks, computing features locally

in MATLAB or through an interface to a mySQL server (enabling distributed computing for large datasets), processing the results of the feature extraction (including options for filtering features on their behavior and feature normalization), and a range of other analytic outputs to facilitate scientific interpretation (including the plots shown in this paper).

Feature Refinement

The feature set used in our previous demonstrations of highly comparative time-series analysis was a largely unrefined set of over 9,600 features (Fulcher et al., 2012, 2013; Fulcher and Jones, 2014). In developing *hctsa*, we made substantial changes to these algorithms in response to a range of robustness and error checks across different time-series data types and time-series lengths. We also addied over 100 new time-series features, removed redundant features (i.e., features with perfectly correlated outputs across a set of 1,000 diverse empirical time series), and improved efficiency of MATLAB code through direct modifications and by utilizing mex functions. All specific changes can be found in the tracked history of *hctsa* at github.com/benfulcher/hctsa/commits/master (NB: *hctsa* development began in 2013). This refinement process yielded a set of 7,749 features that are included in the version of *hctsa* described in this paper. This set will be refined further in future through contributions from the time-series analysis community, as a non-static, living library.

Feature Filtration and Normalization

For any given analysis, we filtered out features that were constant across the dataset or contained any 'special' values (e.g., due to applying a method that is inappropriate for the data, such as fitting a positive-only distribution to data that are not positive only, or attempting to fit a model to the data that does not converge, etc.). Due to this filtering, a different number of total features will be usable for a given dataset, depending on its properties.

When searching for discriminative individual features, we did not normalize or rescale feature values to enable results to be interpreted in the natural scale of each feature. However, when computing the Principal Components of a dataset, or learning a classifier in the full feature space, we normalized each feature to the unit interval using a scaled robust sigmoid function (Fulcher et al., 2013):

$$\tilde{\boldsymbol{f}} = \left[1 + exp\left(-\frac{\boldsymbol{f} - m_{\boldsymbol{f}}}{1.35 r_{\boldsymbol{f}}}\right)\right], \quad \text{(Equation 1)}$$

where $\tilde{\boldsymbol{f}}$ represents the normalized feature values across a time-series dataset, $\boldsymbol{f}$ is the vector of un-normalized feature values, $m_{\boldsymbol{f}}$ is the median of $\boldsymbol{f}$, and $r_{\boldsymbol{f}}$ is its interquartile range.

Classification

For multi-class classification, we trained linear support vector machine classifiers in MATLAB (2015b) (a product of The MathWorks, Natick, MA) using the **fitcecoc** function with a linear kernel SVM. To compare single univariate features, we used simple linear discriminant analysis (using **classify**). When training SVM classifiers, we weighted each observation, x, as the inverse probability of its class label across the dataset to account for class imbalance.

Calculation Time

For the datasets analyzed here, to compute all 7,749 features on a 16-core machine, each *Drosophila* time series (of length 4 320 samples) took 48 s, and each *C. elegans* time series (most are of length 26,196 samples) took 11.14 min. Details on how compute time scales with time-series length, and the mechanisms included in *hctsa* for distributing calculations across cores on a single machine, and/or across a distributed computing platform, are described in the online documentation (www.gitbook.com/book/benfulcher/hctsa-manual).

QUANTIFICATION AND STATISTICAL ANALYSIS

Due to imbalance of observations across the multiclass classification problems investigated here, we report balanced classification accuracy, C_{bal}, over m classes in terms of the number of correctly identified examples of a given class ti, and the total number of examples of each class, ci, as

$$C_{bal} = \frac{1}{m}\sum_{i=1}^{m}\frac{t_i}{c_i} \quad \text{(Equation 2)}$$

Balancing the accuracy in this way ensures that all classes contribute equally to the classification statistic. 10-fold cross-validation is used to prevent overfitting leading to optimistic performance estimates.

Estimating the number of individually significant features for a given outcome follows a permutation test, pooling across features for 50 random permutations. For the example of 7,000 features, this yields a null distribution containing 7,000 x 50 = 350,000 samples, from which p-values are estimated, and then corrected for multiple comparisons using the method of Benjamini and Hochberg using the **mafdr** function in MATLAB.

Detailed steps, including all code for reproducing all figures and anlyses presented here are described in the Data and Software Availability section below. This code relies on the following analysis/visualization functions of *hctsa*: **TS_classify** (classification of labeled groups using the set of all features), **TS_TopFeatures** (determine individual features with significant predictive ability of the class labels), **TS_plot_pca** (to generate plots of time series projected into a low-dimensional feature space), and **TS_SingleFeature** (to plot distributions of individual features).

DATA AND SOFTWARE AVAILABILITY

Dataset Availability

The two datasets analyzed here, including the labeled time series and the full results of *hctsa* feature extraction, are available in the form of MATLAB files (.mat) for *C. elegans*: https://dx.doi.org/10.4225/03/580478f951263, and for *Drosophila*: https://dx.doi.org/10.4225/03/5804798d2a2ec.

Code Availability

Analyses presented here were computed using v0.92 of *hctsa*, which contains a total of 7,749 features. The *hctsa* software is freely available at github.com/benfulcher/hctsa/. A categorized list of all time-series analysis code files included in *hctsa*, with brief descriptions, is provided with the online documentation (www.gitbook.com/book/benfulcher/hctsa-manual) and in the Supplemental Information. Analysis pipelines used to produce the results reported here (as well as many other outputs from *hctsa*) are available at github.com/benfulcher/hctsa_phenotypingWorm/ and github.com/benfulcher/hctsa_phenotypingFly/ for the *C. elegans* and *Drosophila* datasets, respectively.