**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 (Hirpcsak 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
**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
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, \textit{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 \textit{unc-38} (which encodes a nicotinic acetylcholine receptor alpha subunit) and \textit{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-Marín 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 burrly 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).

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:

- **KEY RESOURCES TABLE**
- **CONTACT FOR REAGENT AND RESOURCE SHARING**
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - Caenorhabditis elegans Movement Speed Data
  - Drosophila melanogaster Movement Speed Data
- **METHOD DETAILS**
  - Feature Refinement
  - Feature Filtration and Normalization
  - Classification
  - Calculation Time
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
- **DATA AND SOFTWARE AVAILABILITY**
  - Dataset Availability
  - Code Availability

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 NIMH 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


