TIME SERIES ANALYSIS

Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality

Hao Ye and George Sugihara*

In ecological analysis, complexity has been regarded as an obstacle to overcome. Here we present a straightforward approach for addressing complexity in dynamic interconnected systems. We show that complexity, in the form of multiple interacting components, can actually be an asset for studying natural systems from temporal data. The central idea is that multidimensional time series enable system dynamics to be reconstructed from multiple viewpoints, and these viewpoints can be combined into a single model. We show how our approach, multiview embedding (MVE), can improve forecasts for simulated ecosystems and a mesocosm experiment. By leveraging complexity, MVE is particularly effective for overcoming the limitations of short and noisy time series and should be highly relevant for many areas of science.

Complex interconnected systems pose a major challenge to scientific study in a variety of fields, including ecology, finance, neuroscience, and medicine (*1*–*4*). Although widely used, the common approach of reducing these systems to linearly independent components overlooks important interactions for the sake of computational tractability. Thus, many statistical frameworks (e.g., principal components analysis, generalized linear models, multivariate autoregressive models) assume that causal factors do not interact with each other and have independent or additive effects on a response variable. This simplification can lead to problems in identifying associations (*5*, *6*) or predicting out-of-sample behavior (*7*). Conversely, complex equation-based models that explicitly account for each interaction [e.g., end-to-end ecosystem models (*8*)] have great intuitive appeal but often have too many parameters to be precisely determined given the available data [the "curse of dimensionality" (*9*)], even assuming that the model structure is generally correct. These issues are particularly acute in biological fields where the relevant units (e.g., species or other variables) may not behave according to fundamental equations (*10*) and where data sets are often cross-sectionally wide (e.g., census many interacting species) but short in the time dimension (*11*, *12*).

One solution to the problem of uncertain model structure and unknown model equations is the framework of empirical dynamic modeling (EDM) (*13*–*15*), which uses attractor manifolds reconstructed from time series data to enable the study of systems [see brief introductory animation http://tinyurl.com/EDM-intro (*5*)]. If system behavior is governed by deterministic rules, then attractor manifolds exist and can be built from lags of a single variable (*16*) or multivariately from combinations of variables (*15*, *17*, *18*). However, because these manifolds are empirical, data limitations can be problematic, especially when time series are short and noisy. For example, with short time series, reconstructed attractors may be sparsely populated, which impedes accurate inference of dynamics from nearest neighbors. Furthermore, observational error will result in reduced precision; even when time series are long enough to densely populate the attractor, the nearest neighbors may not form a smooth map.

Here we introduce multiview embedding (MVE) as a general approach that exploits complexity to amplify information and address these issues. The basic idea is straightforward. In interconnected systems with multiple time series observations, many different variable combinations are possible (*16*–*18*) (Fig. 1A). Each reconstruction created from a particular combination of variables can be thought of as a caricature (view) of the system that contains distinctive information when constrained by finite and noisy data (*19*). For example, Fig. 1B show predictions for a three-species food-chain simulation (*20*) using models built on the same 25-point time interval. Here, predictions are produced from univariate models (views using lags of single variables x, y, or z), and each model view exhibits distinct errors. Even with valid embedding coordinates, 25 points may be too few to fully resolve the system behavior—that is, the manifold may be too sparse, especially with observation error. However, because each view is better at resolving different portions of the system, a more complete model should be possible through combination.

A simple and straightforward implementation to combine multiple views is as follows: In contrast to conventional simplex projection (*13*), where a forecast is based on the weighted average of the nearest neighbors in a single view (Fig. 2A), we examine the top k reconstructions, and use the single nearest neighbor from each (Fig. 2B). The MVE forecast (e.g., for variable y) is then defined as a simple average

$$\hat{y}_{t+1} = \frac{1}{k}\sum_{i=1}^{k} y_{nn^i(t)+1}$$

where $nn^i(t)$ is the time index of the nearest neighbor in the ith attractor. In essence, this approach is intended to mitigate prediction errors that occur when nearest neighbors are misidentified or inaccurately weighted based on distance (e.g., due to finite, noisy data). Instead, each of the k neighbors comes from a different view of the system; thus, each corresponding prediction $[y_{nn^i(t)+1}]$ is effectively weighted by how frequently it appears as a nearest neighbor among the top k reconstructions. This is a more robust indication of its true similarity to the target point. This simple implementation of MVE produces forecasts that are substantially better at covering the full range of system behavior than the individual univariate models (Fig. 1, B and C).

The information leverage of MVE follows from two results on dynamic systems arising from Takens' theorem (*16*): Causal effects are recorded in the time series of affected variables (*5*), and each combination of variables and lags is a valid embedding (*17*, *18*). These two properties mean that the interconnectedness of complex systems is actually a boon: Whenever variable x has an influence on some other variable y, information about x is recorded in y and can be recovered. Because each embedding filters this information in a different way, combining multiple models can be highly advantageous (Fig. 1C)—an advantage that increases as the system becomes more complex. In fact, the number of possible reconstructions grows combinatorially with the number of variables. Given l lags for each of n variables, the number of E-dimensional variable combinations is

$$m = \binom{nl}{E} - \binom{n(l-1)}{E}$$

For a simple system with 10 variables (and up to three lags each), the number of distinct three-dimensional combinations is nearly 3000. Although all variable combinations are valid embeddings, with limited data they will not resolve the system equally well. Therefore, we use only the top k reconstructions, as ranked by in-sample forecast accuracy (ρ, correlation between observations and predictions), and apply the heuristic of $k = \sqrt{m}$ (*21*–*23*).

To quantify performance, we compare the out-of-sample forecast skill of this multimodel approach with standard nonlinear methods: a univariate model using only lags of the variable being forecast and a multivariate model defined by the variable combination with the highest in-sample ρ. Figure 3 shows this comparison for three simple

Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive 0202, La Jolla, CA 92093-0202, USA.
*Corresponding author. Email: gsugihara@ucsd.edu

ecosystem simulations with 10% observational error [methodological details in (*24*)]: a three-species coupled logistic, a three-species food chain (*20*), and a three-stage flour beetle model (*25*). In nearly all conditions, MVE produces better forecasts (higher ρ) compared with the univariate and multivariate methods. Results were broadly similar when repeated with a more complex 12-species resource competition system (*26*) (figs. S1 and S2).

As a final test, we repeated the analysis using time series data from an 8-year mesocosm experiment of a plankton community isolated from the Baltic Sea (*27*, *28*). This experimental field system exhibits coupled oscillations between predator and prey species, providing a natural experiment for testing MVE. Here, we focus on a subsystem of two predators (rotifers and calanoid copepods) that consume two prey (picocyanobacteria and nanoflagellates) (Fig. 4A). A causality test (*5*) verifies that both prey affect both predators (Fig. 4B), indicating that prey abundances are informative for predicting predator abundances. Just as with the model systems, the multiview approach outperforms the other methods (Fig. 4C). In all cases, other metrics produce qualitatively identical results (figs. S3 to S6).

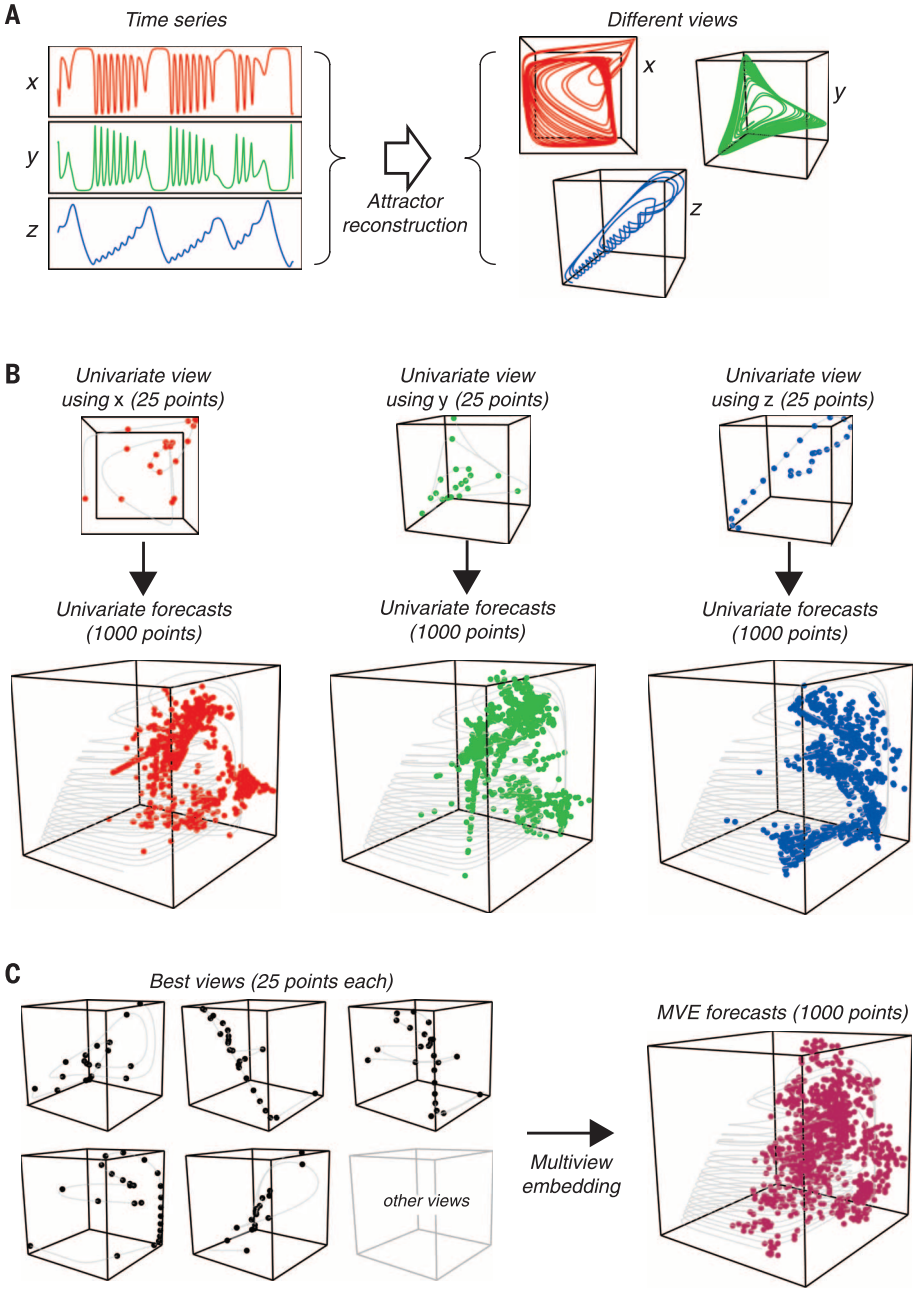

Fig. 1. Schematic for multiview embedding using the three-species food-chain model. (**A**) By combining multiple time series observations of the system, different attractor reconstructions (i.e., views of the system dynamics) are created. Here, the univariate reconstructions using lags of *x* (red), *y* (green), or *z* (blue) are depicted. (**B**) Forecasts based on univariate views of the system (from the same 25 time points of data) give incomplete coverage of the system attractor (gray lines) (*20*). Note that the 1000 predictions (solid points) from each univariate model occupy distinct subsets. (**C**) Combining information from multiple reconstructions [spanning the same 25 time points in (B)], the MVE model gives a clearer depiction of the actual dynamics, resulting in predictions that span more of the original system attractor.

An important concern with any modeling framework is how well it accommodates observational error. For EDM, noise in the data means that reconstructed states of the system are uncertain, affecting all calculations, including the computed distances between states, identification of nearest neighbors, and the final forecast. Depending on the system dynamics and the particular variable combination used for the reconstructed attractor, observational noise can cause large forecast errors (*19*). Our results indicate that as more observational noise is added, forecast skill for all three methods decreases (figs. S7 to S12). However, the use of multiple views in MVE can reduce the effects of noise. Thus, with particularly noisy data, the information advantage of combining multiple views can be more important than selecting a single best model (including the "true" multivariate model composed of the original state variables). This approach to noise reduction builds upon historical approaches in nonlinear state space reconstruction (*19*, *29*) and operates in a way that is fundamentally different from classical frameworks that seek to filter noise by using assumptions about the underlying dynamics and noise structure [e.g., Kalman filters (*30*)].

With longer time series, the single-view multivariate method (using native coordinates) should perform about as well as MVE. With sufficiently long time series, the performance of the two methods is nearly indistinguishable in the absence of observational error (figs. S13 and S14). However, even with small amounts of noise (i.e., 10% added variance), the multivariate approach produces less skillful forecasts than MVE (Fig. 3), suggesting that noise, rather than data length, is the limiting factor. Thus, given the practical constraints of collecting longer time series (true for many natural systems and particularly true for ecosystem studies constrained by short funding horizons), these results show how it can be highly beneficial to combine disparate data sets to leverage signal in synchronous observations.

Nevertheless, it is important that time series be long enough to sufficiently sample the system dynamics. The procedure of selecting the best views can be sensitive to short data segments that are nonrepresentative. For example, the best representation of the system behavior can change over time as dynamics pass through different regimes, such as in our 12-species resource competition model where different groups of species are active at different times (*26*). As a result, with very short temporal data, the multivariate and MVE methods (which rely on in-sample forecast

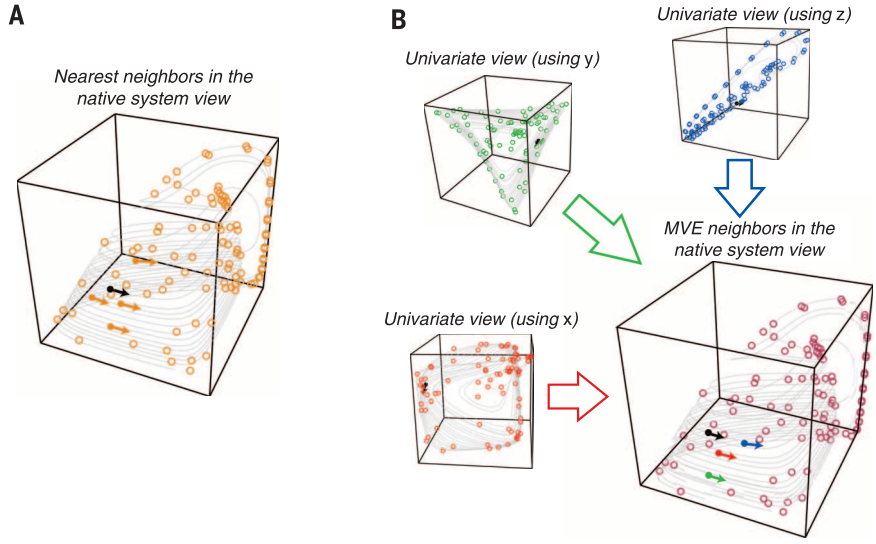

Fig. 2. Nearest-neighbor selection on attractor manifolds. (**A**) In the native system view, the nearest neighbors (solid orange points) to the target point (black) are used to predict the future trajectory. (**B**) MVE selects the single nearest neighbor in each of the different views to produce a more robust model. Here, the nearest neighbors (red, green, and blue) to the target point (black) from the three univariate views (based on lags of *x*, *y*, or *z*, respectively) are used to forecast the future behavior of the target.

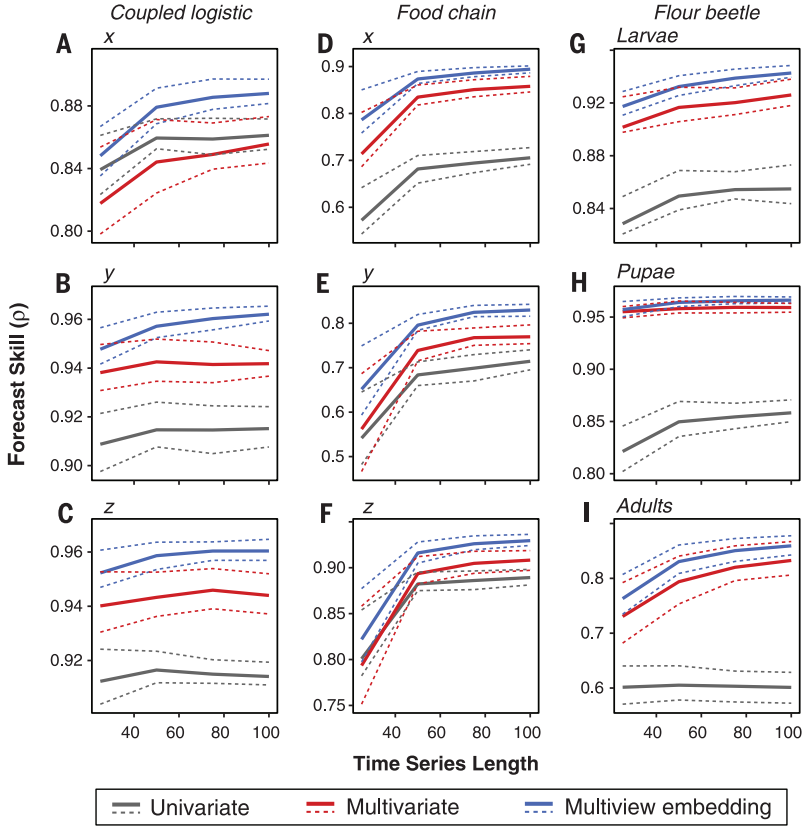

Fig. 3. Comparison of forecast skill for univariate, multivariate, and MVE methods on simulated data with 10% observational error. (**A** to **C**) Forecast skill (ρ, correlation between observations and predictions) versus library size for variables *x*, *y*, and *z* in the three-species coupled logistic. Solid lines indicate average values over 100 randomly sampled libraries; dashed lines denote upper and lower quartiles. (**D** to **F**) Same as (A) to (C) but for the three-species food-chain model (*20*). (**G** to **I**) Same as (A) to (C) but for the flour beetle model (*25*).

skill to select the best views) may show biases [e.g., for N_{11} with 25 data points (figs. S1 and S2)]. Even in simple systems, 25 time points may not provide full coverage of the system (gray areas in Fig. 1C), so it is to be expected that longer time series may be needed if the dynamics are more complex and pass through different regimes.

The example implementation of MVE given here is based on simple model averaging. Whereas this approach has the advantage of transparency and parsimony (involving few parameters), more sophisticated implementations should greatly enhance forecast skill. For example, rather than using the $k = \sqrt{m}$ heuristic (*21*), the optimal number of reconstructions can be tuned. In some cases, forecast skill may be maximized for small k, but in other cases, accuracy continues to increase

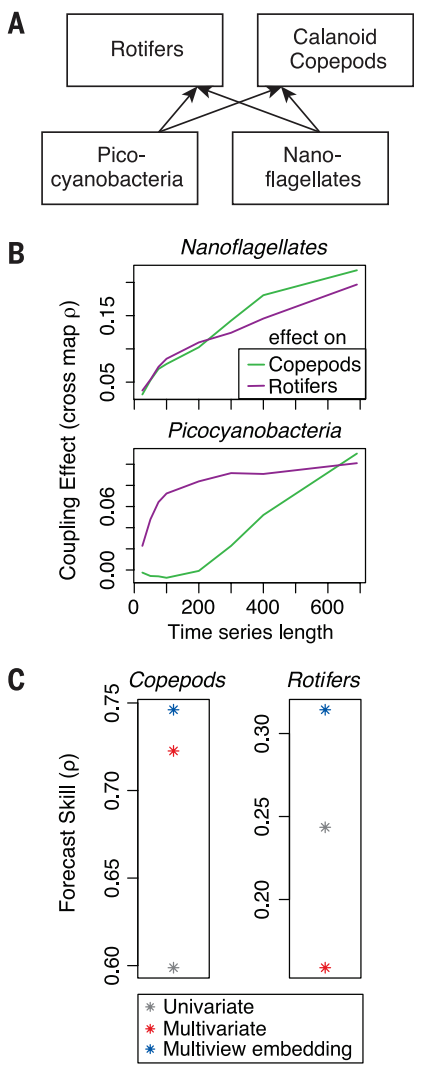

Fig. 4. Analysis of the long-term mesocosm experiment. (**A**) The subsystem examined in this work (*27*, *28*). (**B**) Cross mapping between grazers (calanoid copepods and rotifers) and prey (nanoflagellates and picocyanobacteria) indicates a causal influence of the prey on the grazers. (**C**) Forecast skill (ρ) is higher for MVE than for the univariate or multivariate methods.

until nearly all variable combinations are considered (figs. S15 to S20). Moreover, alternative criteria for selecting candidate views may be desirable to address specific objectives. For example, a more robust but less specifically predictive model could be constructed by selecting variable combinations that are maximally distinct. With enough data, it should even be possible to identify optimal weightings of the different views or have such weightings be state-dependent (e.g., to correct for the state-dependent biases of individual views). Regardless of details, the implementation of MVE demonstrated here is intended to be as simple as possible.

The main innovation of MVE is to leverage the interconnectedness (the shared information) of complex systems. As seen in Fig. 3, improvements in forecast skill can be especially evident for short time series (~25 time points). This result is especially promising given that many current ecological data sets are wide in scope, with many different variables being tracked, but shallow in terms of time series length. Furthermore, the noise-mitigating aspects of MVE are potentially useful for many other applications such as reconstructing historical behavior, signal processing (*31*), or nonlinear system control (*32*). Although the high-dimensionality of complex systems is typically perceived as an obstacle, such complexity is actually an advantage, leading to better clarity and prediction.

REFERENCES AND NOTES

1. R. M. May, S. A. Levin, G. Sugihara, *Nature* **451**, 893–895 (2008).
2. D. Boyd, K. Crawford, *Inf. Commun. Soc.* **15**, 662–679 (2012).
3. J. C. McBride *et al.*, *Neuroimage Clin.* **7**, 258–265 (2014).
4. M. G. M. Olde Rikkert *et al.*, *Crit. Care Med.* **44**, 601–606 (2016).
5. G. Sugihara *et al.*, *Science* **338**, 496–500 (2012).
6. J. Fan, F. Han, H. Liu, *Natl. Sci. Rev.* **1**, 293–314 (2014).
7. D. Lazer, R. Kennedy, G. King, A. Vespignani, *Science* **343**, 1203–1205 (2014).
8. E. A. Fulton, thesis, University of Tasmania (2001).
9. D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," presented at the American Mathematical Society Math Challenges of the 21st Century conference, Los Angeles, CA, 7 to 12 August 2000.
10. D. L. DeAngelis, S. Yurek, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3856–3857 (2015).
11. C.-H. Hsieh, C. Anderson, G. Sugihara, *Am. Nat.* **171**, 71–80 (2008).
12. T. Clark *et al.*, *Ecology* **96**, 1174–1181 (2015).
13. G. Sugihara, R. M. May, *Nature* **344**, 734–741 (1990).
14. G. Sugihara, *Philos. Trans. Phys. Sci. Eng.* **348**, 477–495 (1994).
15. P. A. Dixon, M. J. Milicich, G. Sugihara, *Science* **283**, 1528–1530 (1999).
16. F. Takens, *Dyn. Syst. Turbul. Lect. Notes Math.* **898**, 366–381 (1981).
17. T. Sauer, J. A. Yorke, M. Casdagli, *J. Stat. Phys.* **65**, 579–616 (1991).
18. E. R. Deyle, G. Sugihara, *PLOS ONE* **6**, e18295 (2011).
19. M. Casdagli, S. Eubank, J. D. Farmer, J. Gibson, *Physica D* **51**, 52–98 (1991).
20. A. Hastings, T. Powell, *Ecology* **72**, 896–903 (1991).
21. R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification* (Wiley, 2012).
22. For nearest-neighbor methods, asymptotic convergence requires only that the selection of k satisfies Stone's consistency theorem (*23*). The square root was chosen for simplicity and for its use in the machine learning literature.
23. C. J. Stone, *Ann. Stat.* **5**, 595–620 (1977).
24. See supplementary materials on *Science* Online.
25. B. Dennis, R. A. Desharnais, J. M. Cushing, S. M. Henson, R. F. Costantino, *Ecol. Monogr.* **71**, 277–303 (2001).
26. J. Huisman, F. J. Weissing, *Nature* **402**, 407–410 (1999).
27. R. Heerkloss, G. Klinkenberg, *Verhandlungen - Int. Vereinigung für Theor. und Angew. Limnol.* **26**, 1952–1956 (1998).
28. E. Benincà, K. D. Jöhnk, R. Heerkloss, J. Huisman, *Ecol. Lett.* **12**, 1367–1378 (2009).
29. T. Sauer, *Physica D* **58**, 193–201 (1992).
30. R. E. Kalman, *J. Basic Eng.* **82**, 35–45 (1960).
31. T. L. Carroll, F. J. Rachford, *Chaos* **22**, 023107 (2012).
32. E. Ott, C. Grebogi, J. A. Yorke, *Phys. Rev. Lett.* **64**, 1196–1199 (1990).

ACKNOWLEDGMENTS

We thank S. Glaser, C. Hsieh, E. Deyle, and S. Munch for suggestions and feedback on early drafts of this work.

This work was supported by U.S. Department of Defense Strategic Environmental Research and Development Program 15 RC-2509, Lenfest Ocean Program award 00028335, NSF grant DEB-1020372, the McQuown Chair in the Natural Sciences, and the Sugihara Family Trust. Mesocosm data are available in the appendix of (*28*); model simulation data are available in Data S1.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6302/922/suppl/DC1
Materials and Methods
Figs. S1 to S20
References
Data S1

6 May 2016; accepted 1 August 2016
10.1126/science.aag0863

SINGLE-CELL GENOMICS

Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons

Naomi Habib,[1,2,3]* Yinqing Li,[1,2,3,4]* Matthias Heidenreich,[1,2,3] Lukasz Swiech,[1,2,3] Inbal Avraham-Davidi,[1] John J. Trombetta,[1] Cynthia Hession,[1] Feng Zhang,[1,2,3,5,6]† Aviv Regev[1,7]†

Single-cell RNA sequencing (RNA-Seq) provides rich information about cell types and states. However, it is difficult to capture rare dynamic processes, such as adult neurogenesis, because isolation of rare neurons from adult tissue is challenging and markers for each phase are limited. Here, we develop Div-Seq, which combines scalable single-nucleus RNA-Seq (sNuc-Seq) with pulse labeling of proliferating cells by 5-ethynyl-2′-deoxyuridine (EdU) to profile individual dividing cells. sNuc-Seq and Div-Seq can sensitively identify closely related hippocampal cell types and track transcriptional dynamics of newborn neurons within the adult hippocampal neurogenic niche, respectively. We also apply Div-Seq to identify and profile rare newborn neurons in the adult spinal cord, a noncanonical neurogenic region. sNuc-Seq and Div-Seq open the way for unbiased analysis of diverse complex tissues.

Single-cell RNA sequencing (scRNA-Seq) has extended our understanding of heterogeneous tissues, including the central nervous system (CNS) (*1*–*3*). However, dynamic processes, such as adult neurogenesis, remain challenging to study by scRNA-Seq. First, scRNA-Seq requires enzymatic tissue dissociation (Fig. 1A), which may compromise the integrity of neurons and their RNA content, skew data toward easily dissociated cell types, and is restricted to fetal or young animals (*1*). Second, it is difficult to capture rare cell types, such as adult newborn neurons (*4*), because of limitations in cell tagging and isolation at each phase of the dynamic process.

We therefore developed Div-Seq, a method for RNA-seq of individual, recently divided cells. Div-Seq relies on sNuc-Seq, a single-nucleus isolation and RNA-Seq method compatible with frozen or fixed tissue (Fig. 1A), which enables enrichment of rare labeled cell populations by fluorescence-activated cell sorting (FACS) (fig. S1). Div-Seq combines sNuc-Seq with pulse labeling of dividing cells by 5-ethynyl-2′-deoxyuridine (EdU) (*5*, *6*).

We validated that sNuc-Seq on population of nuclei faithfully represents tissue-level RNA

[1]Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA. [2]Stanley Center for Psychiatric Research, 75 Ames Street, Cambridge, MA 02142, USA. [3]McGovern Institute of Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [6]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [7]Howard Hughes Medical Institute, Koch Institute of Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: zhang@broadinstitute.org (F.Z.); aregev@broadinstitute.org (A.R.)

Harnessing complexity in ecology

Ecology concerns the behavior of complex, dynamic, interconnected systems of populations, communities, and ecosystems over time. Yet ecological time series can be relatively short, owing to practical limits on study duration. Ye and Sugihara introduce an analytical approach called multiview embedding, which harnesses the complexity of short, noisy time series that are common in ecology and other disciplines such as economics. Using examples from published data sets, they show how this approach enhances the tractability of complex data from multiple interacting components and offers a way forward in ecological forecasting.

Science, this issue p. 922