Joshua T. Vogelstein

✉ *jovo@jhu.edu*
jovo.me

I am currently an Assistant Professor of Biomedical Engineering in the Whiting School of Engineering at Johns Hopkins University, where I co-direct the NeuroData lab, whose mission is to fUnderstand and improve animal and machine intelligences worldwide. As of September 2019, according to Google Scholar, I have over 5,000 citations and an h-index of 29.

Our website, neurodata.io, has the most up to date information regarding our team's publications, talks, posters, awards, press, funding, and blog.

Education & Training

08/12 – 08/14 **Senior Research Scientist**, *Dept's of Statistical Sciences & Mathematics & Neurobiology*, Supervised by Mauro Maggioni, Lawrence Carin, Guillermo Sapiro, and David Dunson, Duke University.
Research Big data statistics, network statistics, graph matching.

01/11 – 08/12 **Assistant Research Professor**, *Department of Applied Mathematics and Statistics*, Supervised by Mauro Maggioni, Lawrence Carin, Jon Harer, and David Dunson, Duke University.
Research Big data statistics, network statistics, graph matching.

12/09 – 01/11 **Post-Doctoral Fellow**, *Department of Applied Mathematics and Statistics*, Supervised by Carey E. Priebe, Johns Hopkins University.
Research Statistics of populations of networks.

2003 – 2009 **Ph.D in Neuroscience**,
Johns Hopkins School of Medicine, Supervised by Eric Young,
Dissertation OOPSI: a family of optical spike inference algorithms for inferring neural connectivity from population calcium imaging .

2009 – 2009 **M.S. in Applied Mathematics & Statistics**, *Johns Hopkins University*.

1998 – 2002 **B.A. in Biomedical Engineering**, *Washington University, St. Louis*.

Summer Workshops

06/08 – 07/08 **Molecular Biology Summer Workshop**, *Smith College, Mass, USA*.

07/08 – 07/08 **Advanced Techniques in Molecular Neuroscience**, *Cold Spring Harbor, New York, USA*.

06/05 – 07/05 **Imaging Structure and Function of the Nervous System (audited)**, *Cold Spring Harbor, New York, USA*.

06/04 – 07/04 **Advanced Course in Computational Neuroscience**, *Obidos, Portugal*.

Positions Held

Current Academic Positions

08/14 – now **Assistant Professor**, *Department of Biomedical Engineering*, Johns Hopkins University (JHU).

08/14 – now **Core Faculty**, *Institute for Computational Medicine (ICM)*.

08/14 – now **Core Faculty**, *Center for Imaging Science (CIS)*.

08/15 – now **Steering Committee**, *Kavli Neuroscience Discovery Institute (KNDI)*.

Current Joint Appointments & Affiliations

09/19 – now **Joint Appointment**, *Department of Biostatistics*, Johns Hopkins University (JHU).

08/15 – now **Joint Appointment**, *Department of Applied Mathematics and Statistics*.

08/14 – now **Joint Appointment**, *Department of Neuroscience*.

08/14 – now **Joint Appointment**, *Department of Computer Science*.

08/14 – now **Assistant Research Faculty**, *Human Language Technology Center of Excellence*.
10/12 – now **Affiliated Faculty**, *Institute for Data Intensive Engineering and Sciences*.

Previous Positions & Affiliations

08/14 – 08/18 **Director of Undergraduate Studies**, *Institute for Computational Medicine*.
05/15 – 07/17 **Co-Founder and Faculty Advisor**, *MedHacks*.
10/12 – 08/14 **Endeavor Scientist**, *Child Mind Institute*.
08/12 – 08/14 **Affiliated Faculty**, *Kenan Institute for Ethics*.
Duke University
08/12 – 08/14 **Adjunct Faculty**, *Department of Computer Science*.
07/04 – 07/12 **Chief Data Scientist**, *Global Domain Partners, LLC*.
06/01 – 09/01 **Research Assistant**, *Prof. Randy O'Reilly, Dept. of Psychology*.
University of Colorado
06/00 – 09/00 **Clinical Engineer**, *Johns Hopkins Hospital*.
06/99 – 08/99 **Research Assistant under Dr. Jeffrey Williams**, *Dept. of Neurosurgery, Johns Hopkins Hospital*.
06/98 – 08/98 **Research Assistant under Professor Kathy Cho**, *Dept. of Pathology, Johns Hopkins School of Medicine*.

Industry Work

Board of Directors

- gigantum - d8alab - mind-x - pivotal path

Previous Positions

Awards & Honors

2014 **F1000 Prime Recommended**, *Vogelstein et al. (2014)*.
2013 **Spotlight**, *Neural Information Processing Systems (NIPS)*.
2011 **Trainee Abstract Award**, *Organization for Human Brain Mapping*.
2008 **Spotlight**, *Computational and Systems Neuroscience (CoSyNe)*.
2002 **Dean's List**, *Washington University*.

Peer-Reviewed Journal Publications

(52 articles published/accepted; top 10 cited 2,944 times; H-index 29)

Pre-Prints

Talks

Invited Talks

Other Talks

Posters

Current Funding

5/17 – 4/20 **Multiscale Generalized Correlation: A Unified Distance-Based Correlation Measure for Dependence Discovery**, *NSF*, Shen (PI) 1712947.
7/17 – 6/20 **CRCNS US-German Res Prop: functional computational anatomy of the auditory cortex**, *NIH*, Ratnanather (PI) 1R01DC016784-01.
10/16 – 9/20 **What Would Tukey Do?**, *DARPA D3M*, Priebe (PI) FA8750-17-2-0112.
9/17 – 8/22 **Sensorimotor processing, decision-making, and internal states: towards a realistic multiscale circuit model of the larval zebrafish brain**, *NIH U19*, Engert (PI) 1U19NS104653-01.

1/18 – 12/19 **Connectome Coding at the Synaptic Scale** , *Schmidt Sciences*, Vogelstein (PI).

11/17 – 10/21 **Lifelong Learning Forests**, *DARPA L2M*, Vogelstein (PI).

11/17 – 10/21 **Continual Learning Across Synapses, Circuits, and Brain Areas**, *DARPA L2M*, Tolias (PI).

7/18 – 6/21 **SemiSynBio: Collaborative Research: YeastOns: Neural Networks Implemented in Communication Yeast Cells**, *NSF*, Shulman (PI).

7/17 – 6/19 (Extended) **NeuroNex Innovation Award: Towards Automatic Analysis of Multi-Terabyte Cleared Brains**, *NSF, 16-569 Neural System Cluster*, Vogelstein (PI) 1707298.

Past Funding

10/17 – 9/18 **Brain Ark**, *Dog Star Technologies*, Vogelstein (PI), 90074647.

1/17 – 10/18 **Brain Comp Infra: EAGER: BrainLab CI: Collaborative, Community Experiments with Data-Quality Controls through Continuous Integration**, *NSF*, Burns (PI), ACI-1649880.

5/15 – 8/18 **From RAGs to Riches: Utilizing Richly Attributed Graphs to Reason from Heterogenous Data**, *DARPA*, Vogelstein (PI), N66001-15-C-4041.

9/14 – 6/19 **Synaptomes of Mouse and Man**, *NIH*, Smith (PI), Allen Institute, R01NS092474.

5/14 – 2/16 **Scalable Brain Graph Analyses Using Big-Memory, High-IOPS Compute Architectures**, *DARPA (GRAPHS)*, Burns (PI), DARPA-BAA-13-15.

3/13 – 1/16 **Computational infrastructure for massive neuroscience image stacks**, *NIH/NSF (BIGDATA)*, Mitra (PI), 1R01DA036400.

2/13 – 9/15 **Endeavor Scientists Training Fellowship**, Child Mind Institute, Vogelstein (PI).

9/12 – 8/15 **Data Sharing: The EM Open Connectome Project**, NIH/NIBIB (CRCNS), Burns (PI), 1R01EB016411.

1/14 – 12/14 **Data Readiness Level**, *Laboratory for Analytic Sciences*, Harer (PI).

1/12 – 10/13 **Graph-Based Scalable Analytics for Big Data**, DARPA (XDATA), Andrews (PI), FA8750-12-C-0239.

12/09 – 1/13 **National Center for Applied Neuroscience Project**, *NSF*, RJ Vogelstein (PI).

Mentoring

Post-Doctoral Fellows

08/18 – now **Jesús Arroyo, PhD**, *Post-doctoral Fellow*, CIS, JHU.
Working on graph matching and joint graph embedding.

07/19 – now **Celine Drieu, PhD**, *Post-doctoral Fellow*, Kavli NDI, JHU.
Co-Advised by Assitant Prof. Kuchibhotla, Department of Psychological and Brain Sciences. Working on understanding learning and memory using two-photon calcium imaging.

07/19 – now **Austin Grave, PhD**, *Post-doctoral Fellow*, Kavli NDI, JHU.
Co-Advised by Prof. Richard Huganir, Department of Neuroscience. Working on understanding whole brain synaptic plasticity using genetic engineering and light microscopy imaging.

07/18 – now **Audrey Branch, PhD**, *Post-doctoral Fellow*, Kavli NDI, JHU.
Co-Advised by Prof Michela Gallagher, extending brain clearing experimental technology from mice to rats. Currently with a manuscript on biorxiv.

07/18 – now **Audrey Branch, PhD**, *Post-doctoral Fellow*, Kavli NDI, JHU.
Co-Advised by Prof Michela Gallagher, extending brain clearing experimental technology from mice to rats. Currently with a manuscript on biorxiv.

09/16 – 08/18 **Cencheng Shen, PhD**, *Post-Doctoral Fellow*, CIS, JHU.
Developed Multiscale Graph Correlation, which is currently the premiere hypothesis testing framework, and about to be integrated into SciPy, by far the world's leading scientific computing package. Currently an Assistent Professor in Department of Statistics at University of Delaware, and still an actice collaborator and grantee.

05/16 – 06/17 **Leo Duan, PhD**, *Post-doctoral Fellow*, CIS, JHU.
Went on to do a second postdoc with Leo Dunson (who I did my second postdoc with). Currently an Assistant Professor at University of Florida.

06/16 – 07/17 **Guilherme Franca, PhD**, *Post-doctoral Fellow*, CIS, JHU.
Worked on non-parametric clustering, with an article about to be accepted in PAMI, the leading machine learning journal. Currently a postdoc for Rene Vidal.

PhD Students

08/19 – now **Michael Powell, MSE**, *PhD advisee*, BME, JHU.
Dissertation will focus on explainable artificial intelligence, spearheads collaboration with Andreas Muller, Co-Director of scikit-learn, the world's leading machine learning package.

06/19 – now **Jaewon Chung, MSE**, *PhD advisee*, BME, JHU.
Dissertation will focus on statistics of populations of human networks. Already co-first author and middle author on multiple manuscripts.

08/19 – now **Tommy Athey, BSE**, *PhD advisee*, BME, JHU.
Dissertation will focus on MouseLight project, spearheads collaborations with Prof. Jeremias Sulam and Michael I. Miller.

08/19 – now **Eric Bridgeford, BSE**, *PhD advisee*, Department of Biostatistics, JHU.
Dissertation will focus on statistics of human connectomes and mitigating batch effects. Already first author on several manuscripts under review, and spearheads collaboration with Prof Brian Caffo at Biostatistics.

08/18 – now **Benjamin Pedigo, BSE**, *PhD advisee*, BME, JHU.
Dissertation will focus on analysis and modeling of the world's first whole animal connectome, in collaboration with Marta Zlatic and Albert Cardona (formerly of Janelia Research Campus). Already co-first author and middle author on multiple manuscripts.

08/16 – now **Vikram Chandrashekhar, BSE**, *PhD advisee*, BME, JHU.
Dissertation has focused on extending LDDMM to whole cleared brain datasets, spearheads collaboration with Prof. Karl Deisseroth's lab at Stanford, one of the world's leading neuroscientists.

08/14 – 01/18 **Tyler Tomita, PhD**, BME, JHU.
Developed Sparse Projection Oblique Randomer Forest in his dissertation, currently the best performing machine learning algorithm on a standard suite of over 100 benchmark problems. Currenly a postdoc with Assistant Prof. Chris Honey of Psychology and Brain Sciences.

Masters Students

06/19 – now **Bijan Varjavand**, *MS advisee*, BME, JHU.
Submitted manuscript to PAMI on advancing statistics on populations of networks.

06/19 – now **Sambit Panda**, *MS advisee*, BME, JHU.
Led development of Python implementation of MGC, to be integrated into SciPy.

06/19 – now **Varun Kotharkar**, *MS advisee*, AMS, JHU.
Investigating theoretical advantages of oblique, as compared to axis-aligned, decision trees.

06/18 – now **Drishti Mannan**, *MS advisee*, BME, JHU.
Preparing manuscript introducing novel specification for large attributed networks.

06/18 – 05/19 **Jaewon Chung**, *MSE advisee*, BME, JHU.
Co-first author of manuscript and co-lead developer of Python package for statistical analysis of networks.

08/14 – 06/17 **Greg Kiar, MSE**, BME, JHU.
Lead deveoper of NDMG, the only existing "soup to nuts" pipeline for both functional and diffusion pipelines; co-first author of manuscript under review.

Undergraduate Students

06/19 – now **Ronan Perry**, *BSE*, BME, JHU.

08/14 – 08/18 **Eric Bridgeford, BSE**, BME, JHU.

08/15 – 08/16 **Albert Lee,BSE**, BME, JHU.

06/15 – 12/15 **Ron Boger, BSE**, BME, JHU.

05/15 – 05/16 **Jordan Matelsky, BSE**, CS and Neuroscience, JHU.

02/15 – 05/16 **Ivan Kuznetsov, BSE**, BME, JHU.

Research Assistants

09/19 – now **Ross Lawrence**, *Research Assistant*, BME, JHU.
Responsible for documenting and bug fixing NDMG.

07/19 – now **Ronak Mehta**, *Research Assistant*, BME, JHU.
Finalizing three manuscripts on (1) uncertainty forests, (2) time-series dependence quantification, and (3) lifelong learning forests.

06/19 – now **Devin Crowley**, *Research Assistant*, BME, JHU.
Lead developer of our scalable Python implementaiton of LDDMM.

02/19 – now **Hayden Helm**, *Assistant Research Faculty*, BME, JHU.
Leading research efforts developing theory and methods for lifelong learning.

10/18 – now **Alex Loftus**, *Research Assistant*, BME, JHU.
Current lead developer of NDMG, transitioning from a stand-alone package to be integrated with DiPy.

06/18 – now **Benjamin Falk**, *Research Engineer*, BME, JHU.
Lead software engineer, overseas all development projects, solely responsible for all cloud infrastructure.

06/17 – now **Jesse Patsolic**, *Assistant Research Faculty*, BME, JHU.
Lead developer converting our extensions to decision forests to be merged into sklearn.

Summer Interns

06/19 – 08/19 **Kareef Ullah**, *Summer Intern*, BME, JHU.

06/19 – 08/19 **Shunan Wu**, *Summer Intern*, BME, JHU.

06/19 – 08/19 **Shiyu Sun**, *Summer Intern*, BME, JHU.

06/19 – 08/19 **Sander Shulhoff**, *Summer Intern*, BME, JHU.

06/19 – 08/19 **Kiki Zhang**, *Summer Intern*, BME, JHU.

06/18 – 08/18 **Papa Kobina Van Dyck**, *Summer Intern*, BME, JHU.

Teaching

New Courses Developed

Fall 2019 **NeuroData Design I**, *EN.580.437*, Course Director.
enrollment 46

Spring 2019 **NeuroData Design II**, *EN.580.437*, Course Director.
enrollment 18

Fall 2018 **NeuroData Design I**, *EN.580.437*, Course Director.
enrollment 22

Spring 2017 **NeuroData Design II**, *EN.580.437*, Course Director.
enrollment 14

Fall 2017 **NeuroData Design I**, *EN.580.437*, Course Director.
enrollment 15

Spring 2016 **Upward Spiral of Science**, *EN.580.468*, Course Director.
enrollment 24

Fall 2016 **NeuroData Design I**, *EN.580.437*, Course Director.
enrollment 16

Spring 2015 **Statistical Connectomics**, Course Director.
enrollment 26

Existing Courses Redeveloped

Fall 2015 **Introduction to Computational Medicine**, Co-Teaching, Course Co-Director.

Guest Lectures

Fall 2016 **SBE II**, *EN.580.437*, 2 Lectures.
Fall 2016 **SBE II**, *EN.580.437*, 2 Lectures.
Fall 2016 **SBE II**, *EN.580.437*, 2 Lectures.
Fall 2016 **jason's class**, *EN.580.437*, 2 Lectures.
Fall 2016 **jason's class**, *EN.580.437*, 2 Lectures.
Fall 2016 **will's class**, *EN.580.437*, Guest Lecture.

Educational Workshops

Fall 2016 **dipy's class**, *EN.580.437*, Guest Lecture.
Fall 2016 **berekely's class**, *EN.580.437*, Guest Lecture.

Academic Activities

08/18 – now **Director of Biomedical Data Science Focus Area**.
05/16 – now **Visiting Scientist**, *Howard Hughes Medical Institute*, Janelia Research Campus.
01/11 – now **Co-Founder & Co-Director**, *NeuroData (formerly Open Connectome Project)*.

Commercial Experience

10/18 – now **Advisory Board**, *Mind-X*.
01/17 – now **Co-Founder**, *gigantum*.
01/17 – now **Advisory Board**, *PivotalPath*.
01/16 – now **Co-Founder**, *d8alab*.

Conference and Journal Activities

Reviewer

Annals of Applied Statistics (AOAS), Bioinformatics, Biophysical Journal, IEEE International Conference on eScience, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE Signal Processing Letters, IEEE Transactions on Signal Processing, International Conference on Learning Representations (ICLR), Frontiers in Brain Imaging Methods, Journal of Machine Learning Research (JMLR), Journal of Neurophysiology, Journal of the Royal Statistical Society B (JRSSB), Nature Communications, Nature Methods, Nature Reviews Neuroscience, Network Science, Neural Computation, Neural Information Processing Systems (NIPS), NeuroImage, Neuroinformatics, PLoS One, PLoS Computational Biology, Current Opinion in Neurobiology.

Editorial Board

Guest Associate Editor, *PLoS Computational Biology*.

Editor, *Neurons, Behavior, Data analysis, and Theory*.

Other Activities

Events Organized

Summer 2019 **Organizer**, *NeuroData Workshop*, https://neurodata.devpost.com.
March 2019 **Organizer**, *Neuro Reproducibility Hackashop*, https://brainx3.io/.
Summer 2017 **Organizer**, *NeuroStorm*, https://brainx2.io.
Spring 2016 **Organizer**, *Global Brain Workshop*, http://brainx.io.
Fall 2015 **Co-Organizer**, *BigNeuro2015: Making Sense of Big Neural Data, NIPS Workshop*, http://neurodata.io/bigneuro2015.

Winter 2015 **Organizer**, *Hack@NeuroData*, http://hack.neurodata.io/.

2015 - 2017 **Faculty Superviser**, *MedHacks*, http://medhacks.org/.

Fall 2012 **Co-Organizer**, *Scaling up EM Connectomics Conference*, https://openwiki.janelia.org/wiki/download/attachments/8687459/final+agenda+EM+Connectomics+100512.pdf.

Web Presence and Social Media

Twitter **5,600+ followers**, *https://twitter.com/neuro_data/*, I have had 27.1K impressions in September, 36.5K impressions in August, 37.7K impressions in July, and 32.6K impressions in June..

Website **100,000 visitors**, *https://neurodata.io*.

Languages

Proficient **English, Hebrew, Love, MATLAB, LaTeX**.

Inproficient **R, Python, HTML, CSS**.

Appended Manuscripts

I have appended the most highly cited manuscripts on which I am first author from each academic position (number of citations as of September, 2019):

PhD **JT Vogelstein et al.** , *Fast Nonnegative Deconvolution for Spike Train Inference From Population Calcium Imaging*, Journal of Neurophysiology, 2010.
300 citations

JHU Postdoc **JT Vogelstein et al.** , *The Predictive Capacity of Personal Genome Sequencing*, Science, 2012.
201 citations

Duke Postdoc **JT Vogelstein et al.** , *Discovery of Brainwide Neural-Behavioral Maps via Multiscale Unsupervised Structure Learning*, Science, 2014.
178 citations

JHU Faculty **JT Vogelstein et al.** , *To the Cloud! A Grassroots Proposal to Accelerate Brain Science Discovery*, Neuron, 2016.
23 citations

J Neurophysiol 104: 3691–3704, 2010.
First published June 16, 2010; doi:10.1152/jn.01073.2009.

Fast Nonnegative Deconvolution for Spike Train Inference From Population Calcium Imaging

Joshua T. Vogelstein,[1] Adam M. Packer,[2,3] Timothy A. Machado,[2,3] Tanya Sippy,[2,3] Baktash Babadi,[4] Rafael Yuste,[2,3] and Liam Paninski[4,5]

[1]*Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland;* [2]*Howard Hughes Medical Institute, Chevy Chase, Maryland;* [3]*Department of Biological Sciences,* [4]*Center for Theoretical Neuroscience, and* [5]*Department of Statistics, Columbia University, New York, New York*

Submitted 9 December 2009; accepted in final form 3 June 2010

Vogelstein JT, Packer AM, Machado TA, Sippy T, Babadi B, Yuste R, Paninski L. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J Neurophysiol* 104: 3691–3704, 2010. First published June 16, 2010; doi:10.1152/jn.01073.2009. Fluorescent calcium indicators are becoming increasingly popular as a means for observing the spiking activity of large neuronal populations. Unfortunately, extracting the spike train of each neuron from a raw fluorescence movie is a nontrivial problem. This work presents a fast nonnegative deconvolution filter to infer the approximately most likely spike train of each neuron, given the fluorescence observations. This algorithm outperforms optimal linear deconvolution (Wiener filtering) on both simulated and biological data. The performance gains come from restricting the inferred spike trains to be positive (using an interior-point method), unlike the Wiener filter. The algorithm runs in linear time, and is fast enough that even when simultaneously imaging >100 neurons, inference can be performed on the set of all observed traces faster than real time. Performing optimal spatial filtering on the images further refines the inferred spike train estimates. Importantly, all the parameters required to perform the inference can be estimated using only the fluorescence data, obviating the need to perform joint electrophysiological and imaging calibration experiments.

INTRODUCTION

Simultaneously imaging large populations of neurons using calcium sensors is becoming increasingly popular (Yuste and Katz 1991; Yuste and Konnerth 2005), both in vitro (Ikegaya et al. 2004; Smetters et al. 1999) and in vivo (Göbel and Helmchen 2007; Luo et al. 2008; Nagayama et al. 2007), and will likely continue to improve as the signal-to-noise ratio (SNR) of genetic sensors continues to improve (Garaschuk et al. 2007; Mank et al. 2008; Wallace et al. 2008). Whereas the data from these experiments are movies of time-varying fluorescence intensities, the desired signal consists of spike trains of the observable neurons. Unfortunately, finding the most likely spike train is a challenging computational task, due to limitations on the SNR and temporal resolution, unknown parameters, and analytical intractability.

A number of groups have therefore proposed algorithms to infer spike trains from calcium fluorescence data using very different approaches. Early approaches simply thresholded dF/F [typically defined as $(F - F_b)/F_b$, where F_b is baseline fluorescence; e.g., Mao et al. 2001; Schwartz et al. 1998] to obtain "event onset times." More recently, Greenberg et al. (2008) developed a dynamic programming algorithm to identify individual spikes. Holekamp et al. (2008) then applied an optimal linear deconvolution (i.e., the Wiener filter) to the fluorescence data. This approach is natural from a signal processing standpoint, but does not realize the knowledge that spikes are always positive. Sasaki et al. (2008) proposed using machine learning techniques to build a nonlinear supervised classifier, requiring many hundreds of examples of joint electrophysiological and imaging data to "train" the algorithm to learn what effect spikes have on fluorescence. Vogelstein and colleagues (2009) proposed a biophysical model-based sequential Monte Carlo (SMC) method to efficiently estimate the probability of a spike in each image frame, given the entire fluorescence time series. Although effective, that approach is not suitable for on-line analyses of populations of neurons because the computations run in about real time per neuron (i.e., analyzing 1 min of data requires about 1 min of computational time on a standard laptop computer).

In the present work, a simple model is proposed relating spiking activity to fluorescence traces. Unfortunately, inferring the most likely spike train, given this model, is computationally intractable. Making some reasonable approximations leads to an algorithm that infers the approximately most likely spike train, given the fluorescence data. This algorithm has a few particularly noteworthy features, relative to other approaches. First, spikes are assumed to be positive. This assumption often improves filtering results when the underlying signal has this property (Cunningham et al. 2008; Huys et al. 2006; Lee and Seung 1999; Lin et al. 2004; Markham and Conchello 1999; O'Grady and Pearlmutter 2006; Paninski et al. 2009; Portugal et al. 1994). Second, the algorithm is fast: it can process a calcium trace from 50,000 images in about 1 s on a standard laptop computer. In fact, filtering the signals for an entire population of >100 neurons runs faster than real time. This speed facilitates using this filter on-line, as observations are being collected. In addition to these two features, the model may be generalized in a number of ways, including incorporating spatial filtering of the raw movie, which can improve effective SNR. The utility of the proposed filter is demonstrated on several biological data sets, suggesting that this algorithm is a powerful and robust tool for on-line spike train inference. The code (which is a simple Matlab script) is available for free download from http://www.optophysiology.org.

METHODS

Data-driven generative model

Figure 1 shows data from a typical in vitro epifluorescence experiment (for data collection details see *Experimental methods*

Address for reprint requests and other correspondence: J. T. Vogelstein, Johns Hopkins University, Department of Neuroscience, 3400 N. Charles St., Baltimore, MD 21205 (E-mail: joshuav@jhu.edu).

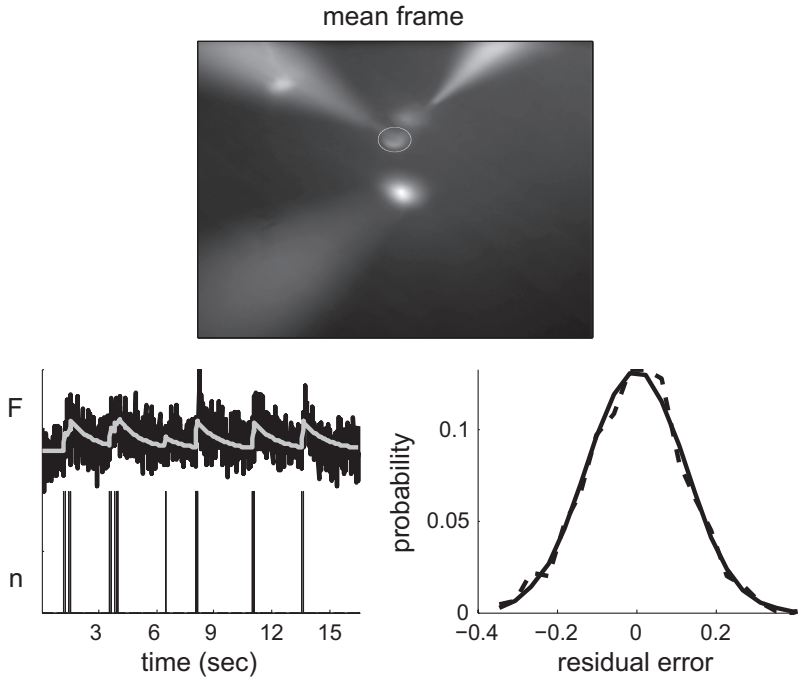

FIG. 1. Typical in vitro data suggest that a reasonable first-order model may be constructed by convolving the spike train with an exponential and adding Gaussian noise. *Top panel*: the average (over frames) of a field of view. *Bottom left*: true spike train recorded via a patch electrode (black bars), convolved with an exponential (gray line), superimposed on the Oregon Green BAPTA 1 (OGB-1) fluorescence trace (black line). Whereas the spike train and fluorescence trace are measured data, the calcium is not directly measured, but rather, inferred. *Bottom right*: a histogram of the residual error between the gray and black lines from the *bottom left panel* (dashed line) and the best-fit Gaussian (solid line). Note that the Gaussian model provides a good fit for the residuals here.

later in this section). The *top panel* shows the mean frame of this movie, including four neurons, three of which are patched. To build the model, the pixels within a region of interest (ROI) are selected (white circle). Given the ROI, all the pixel intensities of each frame can be averaged, to get a one-dimensional fluorescence time series, as shown in the *bottom left panel* (black line). By patching onto this neuron, the spike train can also be directly observed (black bars; *bottom left*). Previous work suggests that this fluorescence signal might be well characterized by convolving the spike train with an exponential and adding noise (Yuste and Konnerth 2005). This model is confirmed by convolving the true spike train with an exponential (gray line; *bottom left*) and then looking at the distribution of the residuals. The *bottom right panel* shows a histogram of the residuals (dashed line) and the best-fit Gaussian distribution (solid line).

The preceding observations may be formalized as follows. Assume there is a one-dimensional fluorescence trace $\boldsymbol{F}$ from a neuron [throughout this text $\boldsymbol{X}$ indicates the vector $(X_1, \ldots, X_T)$, where T is the index of the final frame]. At time t, the fluorescence measurement F_t is a linear-Gaussian function of the intracellular calcium concentration at that time $[Ca^{2+}]_t$:

$$F_t = \alpha[Ca^{2+}]_t + \beta + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

The parameter α absorbs all experimental variables influencing the scale of the signal, including the number of sensors within the cell, photons per calcium ion, amplification of the imaging system, and so on. Similarly, the offset β absorbs, for example, the baseline calcium concentration of the cell, background fluorescence of the fluorophore, and imaging system offset. The noise at each time ε_t is independently and identically distributed according to a normal distribution with zero mean and σ^2 variance, as indicated by the notation $\overset{iid}{\sim} \mathcal{N}(0, 1)$. This noise results from calcium fluctuations independent of spiking activity, fluorescence fluctuations independent of calcium, and other sources of imaging noise.

Then, assuming that the intracellular calcium concentration $[Ca^{2+}]_t$ jumps by A μM after each spike and subsequently decays back down to baseline C_b μM, with time constant τ s, one can write:

$$[Ca^{2+}]_{t+1} = (1 - \Delta/\tau)[Ca^{2+}]_t + (\Delta/\tau)C_b + An_t \tag{2}$$

where Δ is the time step size—which is the frame duration, or 1/(frame rate)—and n_t indicates the number of times the neuron spiked in frame t. Note that because $[Ca^{2+}]_t$ and F_t are linearly related to one another, the fluorescence scale α and calcium scale A are not identifiable. In other words, either can be set to unity without loss of generality because the other can absorb the scale entirely. Similarly, the fluorescence offset β and calcium baseline C_b are not identifiable, so either can be set to zero without loss of generality. Finally, letting $\gamma = (1 - \Delta/\tau)$, *Eq. 2* can be rewritten by replacing $[Ca^{2+}]_t$ with its nondimensionalized counterpart C_t:

$$C_t = \gamma C_{t-1} + n_t. \tag{3}$$

Note that C_t does not refer to absolute intracellular concentration of calcium, but rather, a relative measure (for a more general model see Vogelstein et al. 2009). The gray line in the *bottom left panel* of Fig. 1 corresponds to the putative $\boldsymbol{C}$ of the observed neuron.

To complete the "generative model" (i.e., a model from which simulations can be generated), the distribution from which spikes are sampled must be defined. Perhaps the simplest first-order description of spike trains is that at each time, spikes are sampled according to a Poisson distribution with some rate:

$$n_t \overset{iid}{\sim} \text{Poisson}\ (\lambda\Delta) \tag{4}$$

where $\lambda\Delta$ is the expected firing rate per bin and Δ is included to ensure that the expected firing rate is independent of the frame rate. Thus *Eqs. 1*, *3*, and *4* complete the generative model.

Goal

Given the above model, the goal is to find the maximum a posteriori (MAP) spike train, i.e., the most likely spike train $\hat{\boldsymbol{n}}$, given the fluorescence measurements, $\boldsymbol{F}$:

$$\hat{\boldsymbol{n}} = \underset{n_t \in \mathbb{N}_0 \forall t}{\text{argmax}}\ P[\boldsymbol{n}|\boldsymbol{F}], \tag{5}$$

where $P[\boldsymbol{n}|\boldsymbol{F}]$ is the posterior probability of a spike train $\boldsymbol{n}$, given the fluorescent trace $\boldsymbol{F}$, and n_t is constrained to be an integer $\mathbb{N}_0 = \{0, 1,$

2, . . . } because of the above assumed Poisson distribution. From Bayes' rule, the posterior can be rewritten:

$$P[\boldsymbol{n}|\boldsymbol{F}] = \frac{P[\boldsymbol{n},\boldsymbol{F}]}{P[\boldsymbol{F}]} = \frac{1}{P[\boldsymbol{F}]}P[\boldsymbol{F}|\boldsymbol{n}]P[\boldsymbol{n}], \quad (6)$$

where $P[\boldsymbol{F}]$ is the evidence of the data, $P[\boldsymbol{F}|\boldsymbol{n}]$ is the likelihood of observing a particular fluorescence trace $\boldsymbol{F}$, given the spike train $\boldsymbol{n}$, and $P[\boldsymbol{n}]$ is the prior probability of a spike train. Plugging the far right-hand side of *Eq. 6* into *Eq. 5*, yields:

$$\hat{\boldsymbol{n}} = \underset{n_t\in\mathbb{N}_0\forall t}{\operatorname{argmax}} \frac{1}{P[\boldsymbol{F}]}P[\boldsymbol{F}|\boldsymbol{n}]P[\boldsymbol{n}] = \underset{n_t\in\mathbb{N}_0\forall t}{\operatorname{argmax}} P[\boldsymbol{F}|\boldsymbol{n}]P[\boldsymbol{n}], \quad (7)$$

where the second equality follows because $P[\boldsymbol{F}]$ merely scales the results, but does not change the relative quality of any particular spike train. Note that the prior $P[\boldsymbol{n}]$ acts as a regularizing term, potentially imposing sparseness or smoothness, depending on the assumed distribution (Seeger 2008; Wu et al. 2006). Both $P[\boldsymbol{F}|\boldsymbol{n}]$ and $P[\boldsymbol{n}]$ are available from the preceding model:

$$P[\boldsymbol{F}|\boldsymbol{n}] = P[\boldsymbol{F}|\boldsymbol{C}] = \prod_{t=1}^{T} P[F_t|C_t], \quad (8a)$$

$$P[\boldsymbol{n}] = \prod_{t=1}^{T} P[n_t], \quad (8b)$$

where the first equality in *Eq. 8*a follows because $\boldsymbol{C}$ is deterministic given $\boldsymbol{n}$, and the second equality follows from *Eq. 1*. Further, *Eq. 8*b follows from the Poisson process assumption, *Eq. 4*. Both $P[F_t|C_t]$ and $P[n_t]$ can be written explicitly:

$$P[F_t|C_t] = \mathcal{N}(\alpha C_t + \beta, \sigma^2), \quad (9a)$$

$$P[n_t] = \text{Poisson}\,(\lambda\Delta), \quad (9b)$$

where both equations follow from the preceding model and the Poisson distribution acts as a sparse prior. Now, plugging *Eq. 9* back into *Eq. 8*, and plugging that result into *Eq. 7*, yields:

$$\hat{\boldsymbol{n}} = \underset{n_t\in\mathbb{N}_0\forall t}{\operatorname{argmax}} \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(F_t-\alpha C_t-\beta)^2}{\sigma^2}\right\}\frac{\exp\{-\lambda\Delta\}(\lambda\Delta)^{n_t}}{n_t!} \quad (10a)$$

$$= \underset{n_t\in\mathbb{N}_0\forall t}{\operatorname{argmax}} \sum_{t=1}^{T}\left\{-\frac{1}{2\sigma^2}(F_t-\alpha C_t-\beta)^2 + n_t \ln \lambda\Delta - \ln n_t!\right\}, \quad (10b)$$

where the second equality follows from taking the logarithm of the right-hand side and dropping terms that do not depend on $\boldsymbol{n}$. Unfortunately, solving *Eq. 10*b exactly is analytically intractable because it requires a nonlinear search over an infinite number of possible spike trains. The search space could be restricted by imposing an upper bound k on the number of spikes within a frame. However, in that case, the computational complexity scales *exponentially* with the number of image frames—i.e., the number of computations required would scale with k^T—which for pragmatic reasons is intractable.

Inferring the approximately most likely spike train, given a fluorescence trace

The goal here is to develop an algorithm to efficiently approximate $\hat{\boldsymbol{n}}$, the most likely spike train given the fluorescence trace. Because of the intractability described earlier, one can approximate *Eq. 4* by replacing the Poisson distribution with an exponential distribution of the same mean (note that potentially more accurate approximations are possible, as described in the DISCUSSION). Modifying *Eq. 10* to incorporate this approximation yields:

$$\hat{\boldsymbol{n}} \approx \underset{n_t>0\forall t}{\operatorname{argmax}} \prod_{t=1}^{T}\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(F_t-\alpha C_t-\beta)^2}{\sigma^2}\right\}(\lambda\Delta)\exp\{-n_t\lambda\Delta\}\right] \quad (11a)$$

$$= \underset{n_t>0\forall t}{\operatorname{argmax}} \sum_{t=1}^{T} -\frac{1}{2\sigma^2}(F_t-\alpha C_t-\beta)^2 - n_t\lambda\Delta \quad (11b)$$

where the second equality follows from taking the log of the right-hand side (logarithm is a monotone function and therefore does not change the relative likelihood of particular spike trains) and dropping terms constant in n_t. Note that the constraint on n_t has been relaxed from $n_t \in \mathbb{N}_0$ to $n_t \geq 0$ (since the exponential distribution can yield any nonnegative number). The exponential prior, much like the Poisson prior, imposes a sparsening effect, by penalizing the objective function for large values of n_t. Further, the exponential approximation makes the optimization problem concave in $\boldsymbol{C}$, meaning that any gradient ascent method guarantees achieving the global maximum (because there are no local maxima, other than the single global maximum). To see that *Eq. 11*b is concave in $\boldsymbol{C}$, rearrange *Eq. 3* to obtain $n_t = C_t - \gamma C_{t-1}$, so *Eq. 11*b can be rewritten:

$$\hat{\boldsymbol{C}} = \underset{C_t-\gamma C_{t-1}>0\forall t}{\operatorname{argmax}} \sum_{t=1}^{T} -\frac{1}{2\sigma^2}(F_t-\alpha C_t-\beta)^2 - (C_t-\gamma C_{t-1})\lambda\Delta \quad (12)$$

which is a sum of terms that are concave in $\boldsymbol{C}$, so the whole right-hand side is concave in $\boldsymbol{C}$. Unfortunately, the integer constraint has been lost, i.e., the answer could include "partial" spikes. This disadvantage can be remedied by thresholding (i.e., setting $n_t = 1$ for all n_t greater than some threshold and the rest setting to zero) or by considering the magnitude of a partial spike at time t as a rough indication of the probability of a spike occurring during frame t. Note the relaxation of a difficult discrete optimization problem into an easier continuous problem is a common approximation technique in the machine learning literature (Boyd and Vandenberghe 2004; Paninski et al. 2009). In particular, the exponential distribution is a convenient nonnegative log-concave approximation of the Poisson (see the DISCUSSION for more details).

Although this convex relaxation makes the problem tractable, the "sharp" threshold imposed by the nonnegativity constraint prohibits the use of standard gradient ascent techniques. This may be rectified by using an "interior-point" method (Boyd and Vandenberghe 2004). Interior-point methods solve nondifferentiable problems indirectly by instead solving a series of differentiable subproblems that converge to the solution of the original nondifferentiable problem. In particular, each subproblem within the series drops the sharp threshold and adds a weighted barrier term that approaches $-\infty$ as n_t approaches zero. Iteratively reducing the weight of the barrier term guarantees convergence to the correct solution. Thus the goal is to efficiently solve:

$$\hat{\boldsymbol{C}}_z = \underset{\boldsymbol{C}}{\operatorname{argmax}} \sum_{t=1}^{T}\left[-\frac{1}{2\sigma^2}(F_t-\alpha C_t-\beta)^2 - (C_t-\gamma C_{t-1})\lambda\Delta + z\ln(C_t-\gamma C_{t-1})\right], \quad (13)$$

where $\ln(\cdot)$ is the "barrier term" and z is the weight of the barrier term (note that the constraint has been dropped). Iteratively solving for $\hat{\boldsymbol{C}}_z$ for z going down to nearly zero guarantees convergence to $\hat{\boldsymbol{C}}$ (Boyd and Vendenberghe 2004). The concavity of *Eq. 13* facilitates using any number of techniques guaranteed to find the global maximum. Because the argument of *Eq. 13* is twice analytically differentiable, one can use the Newton–Raphson technique (Press et al. 1992). The special tridiagonal structure of the Hessian enables each Newton–Raphson step to be very efficient (as described below). To proceed, *Eq. 13* is first rewritten in more compact matrix notation. Note that:

$$\boldsymbol{MC} = \begin{bmatrix} -\gamma & 1 & 0 & 0 & \cdots & 0 \\ 0 & -\gamma & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\gamma & 1 & 0 \\ 0 & \cdots & 0 & 0 & -\gamma & 1 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_{T-1} \\ C_T \end{bmatrix} = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_{T-1} \end{bmatrix}, \tag{14}$$

where $\boldsymbol{M} \in \mathbb{R}^{(T-1)\times T}$ is a bidiagonal matrix. Then, letting $\mathbf{1}$ be a $(T-1)$-dimensional column vector, $\boldsymbol{\beta}$ a T-dimensional column vector of β values, and $\boldsymbol{\lambda} = \lambda\boldsymbol{\Delta}\mathbf{1}$ yields the objective function (*Eq. 13*) in more compact matrix notation (note that throughout we will use the subscript $\odot$ to indicate element-wise operations):

$$\hat{\boldsymbol{C}}_z = \underset{\boldsymbol{MC} \geq_{\odot} \mathbf{0}}{\operatorname{argmax}} -\frac{1}{2\sigma^2} \| \boldsymbol{F} - \alpha\boldsymbol{C} - \boldsymbol{\beta} \|_2^2 - (\boldsymbol{MC})^{\mathrm{T}}\boldsymbol{\lambda} + z \ln_{\odot} (\boldsymbol{MC})^{\mathrm{T}}\mathbf{1}, \tag{15}$$

where $\boldsymbol{MC} \geq_{\odot} \mathbf{0}$ indicates an element-wise greater than or equal to zero, $\ln_{\odot}(\cdot)$ indicates an element-wise logarithm, and $\|x\|_2$ is the standard L_2 norm, i.e., $\|x\|_2^2 = \Sigma_i x_i^2$. When using Newton–Raphson to ascend a surface, one iteratively computes both the gradient $\boldsymbol{g}$ (first derivative) and Hessian $\boldsymbol{H}$ (second derivative) of the argument to be maximized, with respect to the variables of interest ($\boldsymbol{C}$ here). Then, the estimate is updated using $\boldsymbol{C}_z \leftarrow \boldsymbol{C}_z + s\boldsymbol{d}$, where s is the step size and $\boldsymbol{d}$ is the step direction obtained by solving $\boldsymbol{Hd} = \boldsymbol{g}$. The gradient and Hessian for this model, with respect to $\boldsymbol{C}$, are given by:

$$\boldsymbol{g} = -\frac{\alpha}{\sigma^2}(\boldsymbol{F} - \alpha\boldsymbol{C} - \boldsymbol{\beta}) + \boldsymbol{M}^T\boldsymbol{\lambda} - z\boldsymbol{M}^T(\boldsymbol{MC})_{\odot}^{-1} \tag{16a}$$

$$\boldsymbol{H} = \frac{\alpha^2}{\sigma^2}\boldsymbol{I} + z\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{MC})_{\odot}^{-2}\boldsymbol{M} \tag{16b}$$

where the exponents on the vector $\boldsymbol{MC}$ indicate element-wise operations. The step size s is found using "backtracking linesearches," which finds the maximal s that increases the posterior and is between 0 and 1 (Press et al. 1992).

Standard implementations of the Newton–Raphson algorithm require inverting the Hessian, i.e., solving $\boldsymbol{d} = \boldsymbol{H}^{-1}\boldsymbol{g}$, a computation that scales *cubically* with T (requires on the order of T^3 operations). Already, this would be a drastic improvement over the most efficient algorithm assuming Poisson spikes, which would require k^T operations (where k is the maximum number of spikes per frame). Here, because $\boldsymbol{M}$ is bidiagonal, the Hessian is tridiagonal, so the solution may be found in about T operations, via standard banded Gaussian elimination techniques (which can be implemented efficiently in Matlab using $\boldsymbol{H}\backslash\boldsymbol{g}$, assuming $\boldsymbol{H}$ is represented as a sparse matrix) (Paninski et al. 2009). In other words, the above approximation and inference algorithm reduces computations from *exponential* to *linear* time. APPENDIX A contains pseudocode for this algorithm, including learning the parameters, as described in the next section. Note that once $\hat{\boldsymbol{C}}$ is obtained, it is a simple linear transformation to obtain $\hat{\boldsymbol{n}}$, the approximate MAP spike train.

Learning the parameters

In practice, the model parameters $\boldsymbol{\theta} = \{\alpha, \beta, \sigma, \gamma, \lambda\}$ tend to be unknown. An algorithm to estimate the most likely parameters $\hat{\boldsymbol{\theta}}$ could proceed as follows: *1*) initialize some estimate of the parameters $\hat{\boldsymbol{\theta}}$, then *2*) recursively compute $\hat{\boldsymbol{n}}$ using those parameters and update $\hat{\boldsymbol{\theta}}$ given the new $\hat{\boldsymbol{n}}$ until some convergence criterion is met. This approach may be thought of as a pseudoexpectation-maximization algorithm (Dempster et al. 1977; Vogelstein et al. 2009). In the following text, details are provided for each step.

INITIALIZING THE PARAMETERS. Because the model introduced earlier is linear, the scale of $\boldsymbol{F}$ relative to $\boldsymbol{n}$ is arbitrary. Therefore before filtering, $\boldsymbol{F}$ is linearly mapped between zero and one, i.e., $\boldsymbol{F} \leftarrow (\boldsymbol{F} - F_{\min})/(F_{\max} - F_{\min})$, where $F_{\min}$ and $F_{\max}$ are the observed minimum and maximum of $\boldsymbol{F}$, respectively. Given this normalization, α is set to one. Because spiking is sparse in many experimental settings, $\boldsymbol{F}$ tends to be around baseline, so β is initialized to be the median of $\boldsymbol{F}$ and σ is initialized as the median absolute deviation of $\boldsymbol{F}$, i.e., $\sigma = \operatorname{median}_t(|F_t - \operatorname{median}_s(F_s)|)/K$, where $\operatorname{median}_i(X_i)$ indicates the median of X with respect to index i and $K = 1.4785$ is the correction factor when using the median absolute deviation as a robust estimator of the SD of a normal distribution. Because in these data the posterior tends to be relatively flat along the γ dimension (i.e., large changes in γ result in relatively small changes in the posterior), estimating γ is difficult. Further, previous work has shown that results are somewhat robust to minor variations in the time constant (Yaksi and Friedrich 2006); therefore γ is initialized at $1 - \Delta/(1\text{ s})$, which is fairly standard (Pologruto et al. 2004). Finally, λ is initialized at 1 Hz, which is between average baseline and evoked spike rate for data of interest.

ESTIMATING THE PARAMETERS GIVEN $\hat{\boldsymbol{n}}$. Ideally, one could integrate out the hidden variables, to find the most likely parameters:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \int P[\boldsymbol{F}, \boldsymbol{C}|\boldsymbol{\theta}]d\boldsymbol{C} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \int P[\boldsymbol{F}|\boldsymbol{C};\boldsymbol{\theta}]P[\boldsymbol{C}|\boldsymbol{\theta}]d\boldsymbol{C}. \tag{17}$$

However, evaluating those integrals is not currently tractable. Therefore *Eq. 17* is approximated by simply maximizing the parameters given the MAP estimate of the hidden variables:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &\approx \underset{\boldsymbol{\theta}}{\operatorname{argmax}} P[\boldsymbol{F}, \hat{\boldsymbol{C}}|\boldsymbol{\theta}] = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} P[\boldsymbol{F}|\hat{\boldsymbol{C}};\boldsymbol{\theta}]P[\hat{\boldsymbol{n}}|\boldsymbol{\theta}] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln P[\boldsymbol{F}|\hat{\boldsymbol{C}};\boldsymbol{\theta}] + \ln P[\hat{\boldsymbol{n}}|\boldsymbol{\theta}], \end{aligned} \tag{18}$$

where $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{n}}$ are determined using the above-described inference algorithm. The approximation in *Eq. 18* is good whenever most of the mass in the integral in *Eq. 18* is around the MAP sequence $\hat{\boldsymbol{C}}$.[1] The argument from the right-hand side of *Eq. 18* may be expanded:

$$\ln P[\boldsymbol{F}|\hat{\boldsymbol{C}};\boldsymbol{\theta}] + \ln P[\hat{\boldsymbol{n}}|\boldsymbol{\theta}] = \sum_{t=1}^{T} \ln P[F_t|\hat{C}_t;\ \alpha, \beta, \sigma] + \sum_{t=1}^{T} \ln P[\hat{n}_t|\lambda]. \tag{19}$$

Note that the right-hand side of *Eq. 19* decouples λ from the other parameters. The maximum likelihood estimate (MLE) for the observation parameters $\{\alpha, \beta, \sigma\}$ is therefore given by:

$$\begin{aligned} \{\hat{\alpha}, \hat{\beta}, \hat{\sigma}\} &= \underset{\alpha,\beta,\sigma>0}{\operatorname{argmax}} \sum_{t=1}^{T} \ln P[F_t|\hat{C}_t;\beta, \sigma] \\ &= \underset{\alpha,\beta,\sigma>0}{\operatorname{argmax}} -\frac{1}{2}(2\pi\sigma^2) - \frac{1}{2}\left(\frac{F_t - \alpha\hat{C}_t - \beta}{\sigma}\right)^2. \end{aligned} \tag{20}$$

Note that a rescaling of α may be offset by a complementary rescaling of $\hat{\boldsymbol{C}}$. Therefore because the scale of $\hat{\boldsymbol{C}}$ is arbitrary (see *Eqs. 2* and *3*), α can be set to one without loss of generality. Plugging $\alpha = 1$ into *Eq. 20* and maximizing with respect to β yields:

$$\hat{\beta} = \underset{\beta > 0}{\operatorname{argmax}} \sum_{t=1}^{T} -(F_t - \hat{C}_t - \beta)^2. \tag{21}$$

Computing the gradient with respect to β, setting the answer to zero, and solving for $\hat{\beta}$ yields $\hat{\beta} = (1/T)\sum_t (F_t - \hat{C}_t)$. Similarly, computing the gradient of *Eq. 20* with respect to σ, setting it to zero, and solving for $\hat{\sigma}$ yields:

[1] *Equation 18* may be considered a crude Laplace approximation (Kass and Raftery 1995).

$$\hat{\sigma} = \sqrt{\frac{1}{T}\sum_t (F_t - \hat{C}_t - \hat{\beta})^2}, \quad (22)$$

which is simply the root-mean-square of the residual error. Finally, the MLE of λ is given by solving:

$$\hat{\lambda} = \underset{\lambda > 0}{\operatorname{argmax}} \sum_t [\ln(\lambda\Delta) - \hat{n}_t\lambda\Delta], \quad (23)$$

which, again, computing the gradient with respect to λ, setting it to zero, and solving for $\hat{\lambda}$, yields $\hat{\lambda} = T/(\Delta \sum_t \hat{n}_t)$, which is the inverse of the inferred average firing rate.

Iterations stop whenever *1*) the iteration number exceeds some upper bound or *2*) the relative change in likelihood does not exceed some lower bound. In practice, parameter estimates tend to converge after several iterations, given the above initializations.

Spatial filtering

In the preceding text, we assumed that the raw movie of fluorescence measurements collected by the experimenter had undergone two stages of preprocessing before filtering. First, the movie was segmented, to determine ROIs, yielding a vector $\vec{F}_t = (F_{1,t}, \ldots, F_{N_p,t})$, which corresponded to the fluorescence intensity at time t for each of the N_p pixels in the ROI (note that we use the $\vec{X}$ throughout to indicate row vectors in space vs. $\boldsymbol{X}$ to indicate column vectors in time). Second, at each time t, that vector was projected into a scalar, yielding F_t, the assumed input to the filter. In this section, the optimal projection is determined by considering a more general model:

$$F_{x,t} = \alpha_x C_t + \beta_x + \sigma\varepsilon_{x,t}, \quad \varepsilon_{x,t} \overset{iid}{\sim} \mathcal{N}(0, 1), \quad (24)$$

where α_x corresponds to the number of photons that are contributed due to calcium fluctuations C_t, and β_x corresponds to the static photon emission at pixel x. Further, the noise is assumed to be both spatially and temporally white, with standard deviation (SD) σ, in each pixel (this assumption can always be approximately accurate by prewhitening; alternately, one could relax the spatial independence by representing joint noise over all pixels with a covariance matrix Σ_t, with arbitrary structure). Performing inference in this more general model proceeds in a nearly identical manner as before. In particular, the maximization, gradient, and Hessian become:

$$\hat{\boldsymbol{C}}_z = \underset{\boldsymbol{MC} \geq_{\odot} \boldsymbol{0}}{\operatorname{argmax}} -\frac{1}{2\sigma^2}\|\vec{\boldsymbol{F}} - \boldsymbol{C}\vec{\alpha} - \boldsymbol{1}_T\vec{\beta}\|_F^2 - (\boldsymbol{MC})^{\mathrm{T}}\boldsymbol{\lambda} + z\ \ln_{\odot}(\boldsymbol{MC})^{\mathrm{T}}\boldsymbol{1} \quad (25)$$

$$\boldsymbol{g} = (\vec{\boldsymbol{F}} - \boldsymbol{C}\vec{\alpha} - \boldsymbol{1}_T\vec{\beta})^{\mathrm{T}}\frac{\vec{\alpha}^{\mathrm{T}}}{\sigma^2} - \boldsymbol{M}^{\mathrm{T}}\lambda + z\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{MC})_{\odot}^{-1} \quad (26)$$

$$\boldsymbol{H} = -\frac{\vec{\alpha}\vec{\alpha}^{\mathrm{T}}}{\sigma^2}\boldsymbol{I} - z\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{MC})_{\odot}^{-2}\boldsymbol{M}, \quad (27)$$

where $\vec{\boldsymbol{F}}$ is an $N_p \times T$ element matrix, $\boldsymbol{1}_T$ is a column vector of ones with length T, $\boldsymbol{I}$ is an $N_p \times N_p$ identity matrix, and $\|x\|_F$ indicates the Frobenius norm, i.e., $\|x\|_F^2 = \sum_{i,j} x_{i,j}^2$, and the exponents and log operator on the vector $\boldsymbol{MC}$ again indicate element-wise operations. Note that to speed up computation, one can first project the background subtracted ($N_c \times T$)-dimensional movie onto the spatial filter $\vec{\alpha}$, yielding a one-dimensional time series $\boldsymbol{F}$, reducing the problem to evaluating a $T \times 1$ vector norm, as in *Eq. 15*.

The parameters $\vec{\alpha}$ and $\vec{\beta}$ tend to be unknown and thus must be estimated from the data. Following the strategy developed in the previous section, we first initialize the parameters. Because each voxel contains some number of fluorophores, which sets both the baseline fluorescence and the fluorescence due to calcium fluctuations, let both the initial spatial filter and initial background be the median image frame, i.e., $\hat{\alpha}_x = \hat{\beta}_x = \operatorname{median}_t(F_{x,t})$. Given these robust initializations, the maximum likelihood estimator for each α_x and β_x is given by:

$$\{\hat{\alpha}_x, \hat{\beta}_x\} = \underset{\alpha_x,\beta_x}{\operatorname{argmax}}\ P[\boldsymbol{F}_x|\hat{\boldsymbol{C}}] \quad (28a)$$

$$= \underset{\alpha_x,\beta_x}{\operatorname{argmax}} \sum_t \ln P[F_{x,t}|\hat{C}_t] \quad (28b)$$

$$= \underset{\alpha_x,\beta_x}{\operatorname{argmax}} \sum_t \left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(F_{x,t} - \alpha_x\hat{C}_t - \beta_x)^2\right\} \quad (28c)$$

$$= \underset{\alpha_x,\beta_x}{\operatorname{argmax}} -\sum_t (F_{x,t} - \alpha_x\hat{C}_t - \beta_x)^2, \quad (28d)$$

where the first equalities follow from *Eq. 1* and the last equality follows from dropping irrelevant constants. Because this is a standard linear regression problem, let $\boldsymbol{A} = [\hat{\boldsymbol{C}}, \boldsymbol{1}_T]^{\mathrm{T}}$ be a $2 \times T$ element matrix and $\boldsymbol{Y}_x = [\alpha_x, \beta_x]^{\mathrm{T}}$ be a 2×1 element column vector. Substituting $\boldsymbol{A}$ and $\boldsymbol{Y}_x$ into *Eq. 28*d yields:

$$\hat{\boldsymbol{Y}}_x = \underset{\boldsymbol{Y}_x}{\operatorname{argmax}} -\|\boldsymbol{F}_x - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}_x\|_2^2, \quad (29)$$

which can be solved by computing the derivative of *Eq. 29* with respect to $\boldsymbol{Y}_x$ and setting to zero, or using Matlab notation: $\hat{\boldsymbol{Y}}_{\boldsymbol{x}} = \boldsymbol{A}\backslash\boldsymbol{F}_{\boldsymbol{x}}$. Note that solving N_p two-dimensional quadratic problems is more efficient than solving a single ($2 \times N_p$)-dimensional quadratic problem. Also note that this approach does not regularize the parameters at all, by smoothing or sparsening, for instance. In the DISCUSSION we propose several avenues for further development, including the elastic net (Zou and Hastie 2005) and simple parametric models of the neuron. As in the scalar F_t case, we iterate estimating the parameters of this model $\boldsymbol{\theta} = \{\vec{\alpha}, \vec{\beta}, \sigma, \gamma, \lambda\}$ and the spike train $\boldsymbol{n}$. Because of the free scale term discussed earlier, the absolute magnitude of $\vec{\alpha}$ is not identifiable. Thus convergence is defined here by the "shape" of the spike train converging, i.e., the norm of the difference between the inferred spike trains from subsequent iterations, both normalized such that $\max(\hat{n}_t) = 1$. In practice, this procedure converged after several iterations.

Overlapping spatial filters

It is not always possible to segment the movie into pixels containing only fluorescence from a single neuron. Therefore the above-cited model can be generalized to incorporate multiple neurons within an ROI. Specifically, letting the superscript i index the N_c neurons in this ROI yields:

$$\vec{F}_t = \sum_{i=1}^{N_c} \vec{\alpha}^i C_i^i + \vec{\beta} + \vec{\varepsilon}_t, \quad \vec{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 I) \quad (30)$$

$$C_t^i = \gamma^i C_{t-1}^i + n_t^i, \quad n_t^i \overset{iid}{\sim} \text{Poisson}(n_t^i; \lambda_i\Delta) \quad (31)$$

where each neuron is implicitly assumed to be independent and each pixel is conditionally independent and identically distributed with variance σ^2, given the underlying calcium signals. To perform inference in this more general model, let $\boldsymbol{n}_t = [n_t^1, \ldots, n_t^{N_c}]$ and $\boldsymbol{C}_t = [C_t^1, \ldots, C_t^{N_c}]$ be N_c-dimensional column vectors. Then, let $\Gamma = \text{diag}(\gamma^1, \ldots, \gamma^{N_c})$ be an $N_c \times N_c$ diagonal matrix and let $\boldsymbol{I}$ and $\boldsymbol{0}$ be an identity and zero matrix of the same size, respectively, yielding:

$$\boldsymbol{MC} = \begin{bmatrix} -\Gamma & \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & -\Gamma & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & -\Gamma & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\Gamma & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{C}_1 \\ \boldsymbol{C}_2 \\ \vdots \\ \boldsymbol{C}_{T-1} \\ \boldsymbol{C}_T \end{bmatrix} = \begin{bmatrix} \boldsymbol{n}_1 \\ \boldsymbol{n}_2 \\ \vdots \\ \boldsymbol{n}_{T-1} \end{bmatrix} \tag{32}$$

and proceed as before. Note that *Eq. 32* is very similar to *Eq. 14*, except that $\boldsymbol{M}$ is no longer bidiagonal, but rather, block bidiagonal (and $\boldsymbol{C}_t$ and $\boldsymbol{n}_t$ are vectors instead of scalars), making the Hessian block-tridiagonal. Importantly, the Thomas algorithm, which is a simplified form of Gaussian elimination, finds the solution to linear equations with block-tridiagonal matrices in linear time, so the efficiency gained from using the tridiagonal structure is maintained for this block-tridiagonal structure (Press et al. 1992). Performing inference in this more general model proceeds similarly as before, letting $\vec{\boldsymbol{\alpha}} = [\vec{\alpha}^1, \ldots, \vec{\alpha}^{N_c}]$:

$$\widehat{\boldsymbol{C}}_z = \underset{\boldsymbol{MC} \geq_{\odot} \mathbf{0}}{\operatorname{argmax}} -\frac{1}{2\sigma^2} \| \vec{\boldsymbol{F}} - \boldsymbol{C}\vec{\boldsymbol{\alpha}} - \mathbf{1}_T\vec{\beta} \|_F^2 - (\boldsymbol{MC})^{\mathrm{T}}\boldsymbol{\lambda} + z \ln_{\odot} (\boldsymbol{MC})^{\mathrm{T}}\mathbf{1}, \tag{33}$$

$$\boldsymbol{g} = (\vec{\boldsymbol{F}} - \boldsymbol{C}\vec{\boldsymbol{\alpha}} - \mathbf{1}_T\vec{\beta})^{\mathrm{T}} \frac{\vec{\boldsymbol{\alpha}}^{\mathrm{T}}}{\sigma^2} - \boldsymbol{M}^{\mathrm{T}}\lambda + z\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{MC})_{\odot}^{-\mathbf{1}} \tag{34}$$

$$\boldsymbol{H} = -\frac{\vec{\boldsymbol{\alpha}}\vec{\boldsymbol{\alpha}}^{\mathrm{T}}}{\sigma^2}\mathrm{I} - z\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{MC})_{\odot}^{-2}\boldsymbol{M}. \tag{35}$$

If the parameters are unknown, they must be estimated. Initialize $\vec{\beta}$ as above. Then, define $\boldsymbol{\alpha}_x = [\alpha_x^1, \ldots, \alpha_x^{N_c}]^{\mathrm{T}}$ and initialize manually by assigning some pixels to each neuron (of course, more sophisticated algorithms could be used, as described in the DISCUSSION). Given this initialization, iterations and stopping criteria proceed as before, with the minor modification of incorporating multiple spatial filters, yielding:

$$\left\{\widehat{\boldsymbol{\alpha}}_x, \widehat{\beta}_x\right\} = \underset{\alpha_x, \beta_x}{\operatorname{argmax}} -\frac{1}{2}\sum_t (F_{x,t} - \sum_{i=1}^{N_c} \alpha_x^i \widehat{C}_t^i - \beta_x)^2, \tag{36}$$

Now, generalizing the above single spatial filter case, let $\boldsymbol{A} = [\hat{\boldsymbol{C}}, \mathbf{1}_T]^{\mathrm{T}}$ be an $(N_c + 1) \times T$ element matrix and $\boldsymbol{Y}_x = [\boldsymbol{\alpha}_x, \beta_x]^{\mathrm{T}}$ be an $(N_c + 1)$-dimensional column vector. Then, one can again use *Eq. 29* to solve to for $\hat{\boldsymbol{\alpha}}_{\mathbf{x}}$ and $\hat{\beta}_x$ for all x.

Experimental methods

SLICE PREPARATION AND IMAGING. All animal handling and experimentation were done according to the National Institutes of Health and local Institutional Animal Care and Use Committee guidelines. Somatosensory thalamocortical or coronal slices 350–400 μm thick were prepared from C57BL/6 mice at age P14 as described (MacLean et al. 2005). Pyramidal neurons from layer V somatosensory cortex were filled with 50 μM Oregon Green BAPTA 1 hexapotassium salt (OGB-1; Invitrogen, Carlsbad, CA) through the recording pipette or bulk loaded with an acetoxymethyl ester of Fura-2 (Fura-2 AM; Invitrogen). The pipette solution contained 130 mM K-methylsulfate, 2 mM $MgCl_2$, 0.6 mM EGTA, 10 mM HEPES, 4 mM ATP-Mg, and 0.3 mM GTP-Tris (pH 7.2, 295 mOsm). After cells were fully loaded with dye, imaging was performed in one of two ways. First, when using Fura-2, images were collected using a modified BX50-WI upright microscope (Olympus, Melville, NY) with a confocal spinning disk (Solamere Technology Group, Salt Lake City, UT) and an Orca charge-coupled device (CCD) camera from Hamamatsu Photonics (Shizuoka, Japan), at 33 Hz. Second, when using Oregon Green, images were collected using epifluorescence with the C9100-12 CCD camera from Hamamatsu Photonics, with arc-lamp illumination with excitation and emission band-pass filters at 480–500 and 510–550 nm, respectively (Chroma, Rockingham, VT). Images were saved and analyzed using custom software written in Matlab (The MathWorks, Natick, MA).

ELECTROPHYSIOLOGY. All recordings were made using the Multiclamp 700B amplifier (Molecular Devices, Sunnyvale, CA), digitized with National Instruments 6259 multichannel cards and recorded using custom software written using the LabVIEW platform (National Instruments, Austin, TX). Square pulses of sufficient amplitude to yield the desired number of action potentials were given as current commands to the amplifier using the LabVIEW and National Instruments system.

FLUORESCENCE PREPROCESSING. Traces were extracted using custom Matlab scripts to segment the mean image into ROIs. The Fura-2 fluorescence traces were inverted. Because some slow drift was sometimes present in the traces, each trace was Fourier transformed, and all frequencies <0.5 Hz were set to zero (0.5 Hz was chosen by eye); the resulting fluorescence trace was then normalized to be between zero and one.

RESULTS

Main result

The main result of this study is that the fast filter can find the approximately most likely spike train $\hat{\boldsymbol{n}}$, very efficiently, and that this approach yields more accurate spike train estimates than optimal linear deconvolution. Figure 2 depicts a simulation showing this result. Clearly, the fast filter's inferred "spike train" (*third panel*) more closely resembles the true spike train (*second panel*) than the optimal linear deconvolution's inferred spike train (*bottom panel*; Wiener filter). Note that neither filter results in an integer sequence, but rather, each infers a real number at each time.

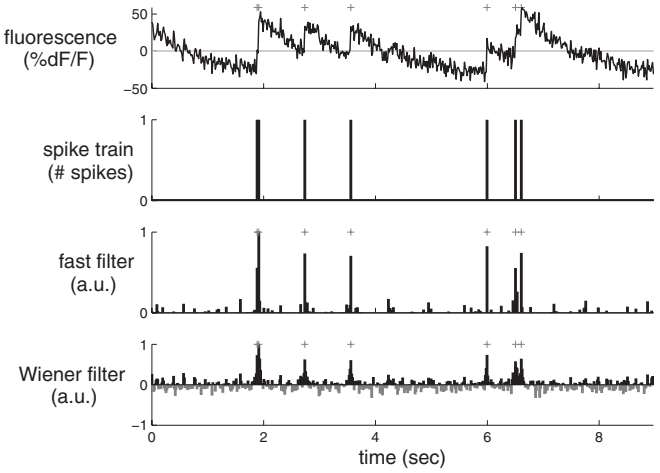

FIG. 2. A simulation showing that the fast filter's inferred spike train is significantly more accurate than the output of the optimal linear deconvolution (Wiener filter). Note that neither filter constrains the inference to be a sequence of integers; rather, the fast filter relaxes the constraint to allow all nonnegative numbers and the Wiener filter allows for all real numbers. The restriction of the fast filter to exclude negative numbers eliminates the ringing effect seen in the Wiener filter output, resulting in a much cleaner inference. Note that the magnitude of the inferred spikes in the fast filter output is proportional to the inferred calcium jump size. *Top panel*: fluorescence trace. *Second panel*: spike train. *Third panel*: fast filter inference. *Bottom panel*: Wiener filter inference. Note that the gray bars in the *bottom panel* indicate negative spikes. Gray + symbols indicate true spike times. Simulation details: $T = 400$ time steps, $\Delta = 33.3$ ms, $\alpha = 1$, $\beta = 0$, $\sigma = 0.2$, $\tau = 1$ s, $\lambda = 1$ Hz. Parameters and conventions are consistent across figures, unless indicated otherwise.

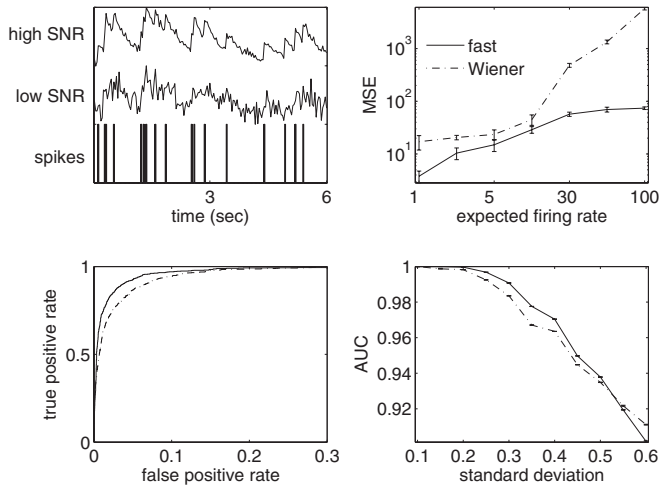

FIG. 3. In simulations, the fast filter quantitatively and significantly achieves higher accuracy than that of the Wiener filter. *Top left*: a spike train (*bottom*) and 2 simulated fluorescence traces, using the same spike train, one with low signal-to-noise ratio (SNR) (*middle*) and one with high SNR (*top*). Simulation parameters: $\tau = 0.5$ s, $\lambda = 3$ Hz, $\Delta = 1/30$ s, $\sigma = 0.6$ (low SNR) and 0.1 (high SNR). Simulation parameters in other panels are the same, except where explicitly noted. *Top right*: mean-squared-error (MSE) for the fast (solid line) and Wiener (dashed-dotted line) filter, for varying the expected firing rate λ. Note that both axes are on a log-scale. Further note that the fast filter has a better (lower) MSE for all expected firing rates. Error bars show SD over 10 repeats. Simulation parameters: $\sigma = 0.2$, $T = 1{,}000$ time steps. *Bottom left*: receiver-operator-characteristic (ROC) curve comparing the fast (solid line) and Wiener (dashed-dotted line) filter. Note that for any given threshold, the Wiener filter has a better (higher) ratio of true positive rate to false positive rate. Simulation parameters as in *top right panel*, except $\sigma = 0.35$ and $T = 10{,}000$ time steps. *Bottom right*: area under the curve (AUC) for fast (solid line) and Wiener (dashed-dotted line) filter as a function of SD (σ). Note that the fast filter has a better (higher) AUC for all σ values until noise gets very high. The 2 simulated fluorescence traces in the *top left panel* show the bounds for SD here. Error bars show SD over 10 repeats.

The Wiener filter implicitly approximates the Poisson spike rate with a Gaussian spike rate (see APPENDIX B for details). A Poisson spike rate indicates that in each frame, the number of possible spikes is an integer, e.g., 0, 1, 2,.... The Gaussian approximation, however, allows any real number of spikes in each frame, including both partial spikes (e.g., 1.4) and *negative* spikes (e.g., −0.8). Although a Gaussian well approximates a Poisson distribution when rates are about 10 spikes per frame, this example is very far from that regime, so the Gaussian approximation performs relatively poorly. Further, the Wiener filter exhibits a "ringing" effect. Whenever fluorescence drops rapidly, the most likely underlying spiking signal is a proportional drop. Because the Wiener filter does not impose a nonnegative constraint on the underlying spiking signal, it infers such a drop, even when it causes n_t to go below zero. After such a drop has been inferred, since no corresponding drop occurred in the true underlying signal here, a complementary jump is often then inferred, to realign the inferred signal with the observations. This oscillatory behavior results in poor inference quality. The nonnegative constraint imposed by the fast filter prevents this because the underlying signal never drops below zero, so the complementary jump never occurs either.

The inferred "spikes," however, are still not binary events when using the fast filter. This is a by-product of approximating the Poisson distribution on spikes with an exponential (cf. *Eq. 11*a) because the exponential is a continuous distribution, versus the Poisson, which is discrete. The height of each spike is therefore proportional to the inferred calcium jump size and can be thought of as a proxy for the confidence with which the algorithm believes a spike occurred. Importantly, by using the Gaussian elimination and interior-point methods, as described in METHODS, the computational complexity of the fast filter is the same as an efficient implementation of the Wiener filter. Note that whereas the Gaussian approximation imposes a shrinkage prior on the inferred spike trains (Wu et al. 2006), the exponential approximation imposes a sparse prior on the inferred spike trains (Seeger 2008).

Figure 3 quantifies the relative performance of the fast and Wiener filters. The *top left panel* shows a typical simulated spike train (*bottom*), a corresponding relatively low SNR fluorescence trace (*middle*), and a relatively high SNR fluorescence trace (*top*), as examples. The *top right panel* compares the mean-squared-error (MSE) of the inferred spike trains using the fast (solid) and Wiener (dashed) filters, as a function of expected firing rate. Clearly, the fast filter has a better (lower) MSE for all rates. The *bottom left panel* shows a receiver-operator-characteristic (ROC) curve (Green and Swets 1966) for another simulation. Again, the fast filter dominates the Wiener filter, having a higher true positive rate for every false negative rate. Finally, the *bottom right panel* shows that the area under the curve (AUC) of the fast filter is better (higher) than that of the Wiener filter until the noise is very large. Collectively, these analyses suggest that for a wide range of firing rates and signal quality, the fast filter outperforms the Wiener filter.

Although in Fig. 2 the model parameters were provided, in the general case, the parameters are unknown and must therefore be estimated from the observations (as described in *Learning the parameters* in METHODS). Importantly, this algorithm does not require labeled training data, i.e., there is no need for joint imaging and electrophysiological experiments to estimate the parameters governing the relationship between the two. Figure 4 shows another simulated example; in this example, however, the parameters are estimated from the observed

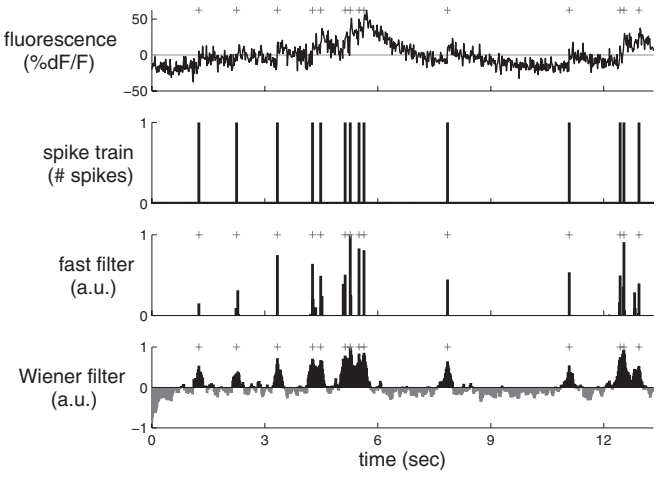

FIG. 4. A simulation showing that the fast filter achieves significantly more accurate inference than that of the Wiener filter, even when the parameters are unknown. For both filters, the appropriate parameters were estimated using only the data shown above, unlike Fig. 2, in which the true parameters were provided to the filters. Simulation details different from those in Fig. 2: $T = 1{,}000$ time steps, $\Delta = 16.7$ ms, $\sigma = 0.4$.

fluorescence trace. Again, it is clear that the fast filter far outperforms the Wiener filter.

Given the preceding two results, the fast filter was applied to biological data. More specifically, by jointly recording electrophysiologically and imaging, the true spike times are known and the accuracy of the two filters can be compared. Figure 5 shows a result typical of the 12 joint electrophysiological and imaging experiments conducted (see METHODS for details). As in the simulated data, the fast filter output is much "cleaner" than the Wiener filter: spikes are more well defined, and not spread out, due to the sparse prior imposed by the exponential approximation. Note that this trace is typical of epifluorescence techniques, which makes resolving individual spikes quite difficult, as evidenced by a few false positives in the fast filter. Regardless, the fast filter output is still more accurate than the Wiener filter, both as determined qualitatively by eye and as quantified (described in the following text). Furthermore, although it is difficult to see in this figure, the first four events are actually pairs of spikes, reflected by the width and height of the corresponding inferred spikes when using the fast filter. This suggests that although the scale of $\boldsymbol{n}$ is arbitrary, the fast filter can correctly ascertain the number of spikes within spike events.

Figure 6 further evaluates this claim. While recording and imaging, the cell was forced to spike once, twice, or thrice for each spiking event. The fast filter infers the correct number of spikes in each event. On the contrary, there is no obvious way to count the number of spikes within each event when using the Wiener filter. We confirm this impression by computing the correlation coefficient, r^2, between the sum of each filter's output and the true number of spikes, for all 12 joint electrophysiological and imaging traces. Indeed, whereas the fast filter's r^2 was 0.47, the Wiener filter's r^2 was -0.01 (after thresholding all negative spikes), confirming that the Wiener filter output cannot reliably convey the number of spikes in a fluorescence trace, whereas the fast filter can. Furthermore, varying the magnitude of the threshold for the Wiener filter to discard more "low-amplitude noise" could increase the magnitude of r^2 (≤ 0.24), still significantly lower than the fast filter's r^2 value. On the other hand, no amount of thresholding

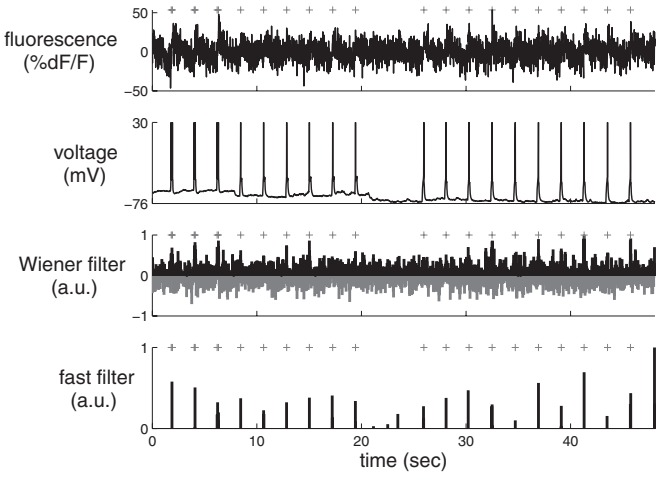

FIG. 5. In vitro data showing that the fast filter significantly outperforms the Wiener filter, using OGB-1. Note that all the parameters for both filters were estimated only from the fluorescence data in the *top panel* (i.e., not considering the voltage data at all). + symbols denote true spike times extracted from the patch data, not inferred spike times from $\boldsymbol{F}$.

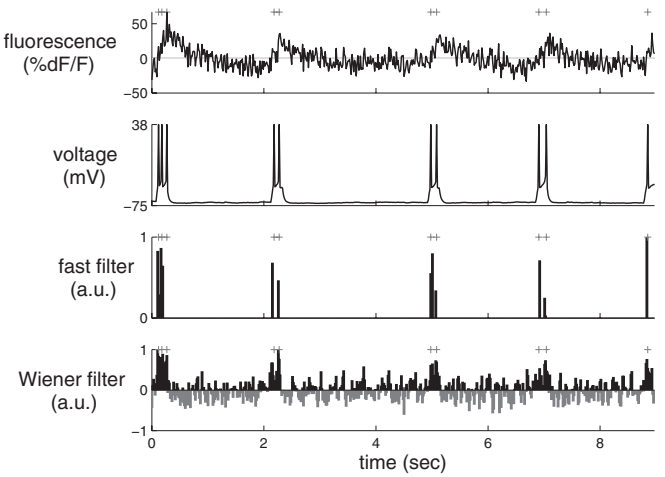

FIG. 6. In vitro data with multispike events, showing that the fast filter can often resolve the correct number of spikes within each spiking event, while imaging using OGB-1, given sufficiently high SNR. It is difficult, if not impossible, to count the number of spikes given the Wiener filter output. Recording and fitting parameters as in Fig. 5. Note that the parameters were estimated using a 60-s-long recording, of which only a fraction is shown here, to more clearly depict the number of spikes per event.

the fast filter yielded an improved r^2, indicating that thresholding the output of the fast filter is unlikely to improve spike inference quality.

On-line analysis of spike trains using the fast filter

A central aim for this work was the development of an algorithm that infers spikes fast enough to use on-line while imaging a large population of neurons (e.g., >100). Figure 7 shows a segment of the results of running the fast filter on 136 neurons, recorded simultaneously, as described earlier in *Experimental methods*. Note that the filtered fluorescence signals show fluctuations in spiking much more clearly than the unfiltered fluorescence trace. These spike trains were inferred in less than imaging time, meaning that one could infer spike trains for the past experiment while conducting the subsequent experiment. More specifically, a movie with 5,000 frames of 100 neurons can be analyzed in about 10 s on a standard desktop computer. Thus if that movie was recorded at 50 Hz, whereas collecting the data would require 100 s, inferring spikes would require only 10 s, a 10-fold improvement over real time.

Extensions

Earlier in METHODS, *Data-driven generative model* describes a simple principled first-order model relating the spike trains to the fluorescence trace. A number of the simplifying assumptions can be straightforwardly relaxed, as described next.

Replacing Gaussian observations with poisson. In the preceding text, observations were assumed to have a Gaussian distribution. The statistics of photon emission and counting, however, suggest that a Poisson distribution would be more natural in some conditions, especially for two-photon data (Sjulson and Miesenböck 2007), yielding:

$$F_t \overset{iid}{\sim} \text{Poisson} \quad (\alpha C_t + \beta), \tag{37}$$

where $\alpha C_t + \beta \geq 0$. One additional advantage to this model over the Gaussian model is that the variance parameter σ^2 no

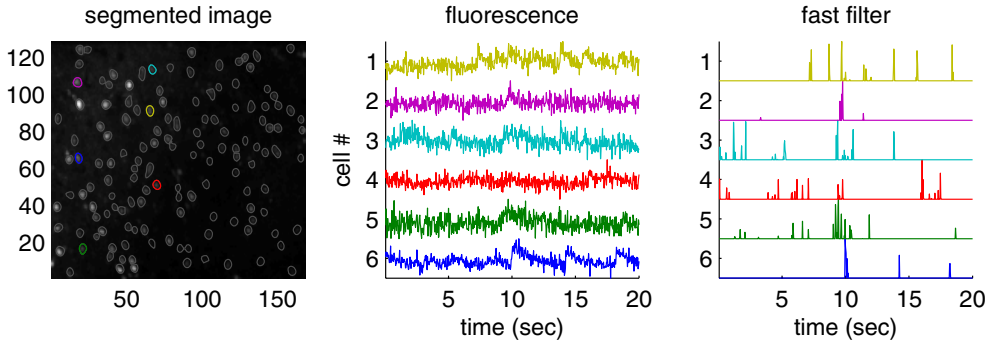

FIG. 7. The fast filter infers spike trains from a large population of neurons imaged simultaneously in vitro, faster than real time. Specifically, inferring the spike trains from this 400-s-long movie including 136 neurons required only about 40 s on a standard laptop computer. The inferred spike trains much more clearly convey neural activity than the raw fluorescence traces. Although no intracellular "ground truth" is available from these population data, the noise seems to be reduced, consistent with the other examples with ground truth. *Left*: mean image field, automatically segmented into regions of interest (ROIs), each containing a single neuron using custom software. *Middle*: example fluorescence traces. *Right*: fast filter output corresponding to each associated trace. Note that neuron identity is indicated by color across the 3 panels. Data were collected using a confocal microscope and Fura-2, as described in METHODS.

longer exists, which might make learning the parameters simpler. Importantly, the log-posterior is still concave in $\boldsymbol{C}$, as the prior remains unchanged, and the new log-likelihood term is a sum of terms concave in $\boldsymbol{C}$:

$$\ln P[\boldsymbol{F}|\boldsymbol{C}] = \sum_{t=1}^{T} \ln P[F_t|C_t] = \sum_{t=1}^{T} \{F_t \ln(\alpha C_t + \beta) - (\alpha C_t + \beta) - \ln(F_t!)\}. \quad (38)$$

The gradient and Hessian of the log-posterior can therefore be computed analytically by substituting the above likelihood terms for those implied by *Eq. 1*. In practice, however, modifying the filter for this model extension did not seem to significantly improve inference results in any simulations or data available at this time (not shown).

Allowing for a time-varying prior. In *Eq. 4*, the rate of spiking is a constant. Often, additional knowledge about the experiment, including external stimuli or other neurons spiking, can provide strong time-varying prior information (Vogelstein et al. 2009). A simple model modification can incorporate that feature:

$$n_t \overset{iid}{\sim} \text{Poisson}(\lambda_t \Delta), \quad (39)$$

where λ_t is now a function of time. Approximating this time-varying Poisson with a time-varying exponential with the same time-varying mean (similar to *Eq. 11*a) and letting $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_T]^{\mathrm{T}}\Delta$, yields an objective function very similar to *Eq. 15*, so log-concavity is maintained and the same techniques may be applied. However, as before, this model extension did not yield any significantly improved filtering results (not shown).

Saturating fluorescence. Although all the abovementioned models assumed a *linear* relationship between F_t and C_t, the relationship between fluorescence and calcium is often better approximated by the nonlinear Hill equation (Pologruto et al. 2004). Modifying *Eq. 1* to reflect this change yields:

$$F_t = \alpha \frac{C_t}{C_t + k_d} + \beta + \varepsilon_t \qquad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \quad (40)$$

Importantly, log-concavity of the posterior is no longer guaranteed in this nonlinear model, meaning that converging to the global maximum is no longer guaranteed. Assuming a good initialization can be found, however, and *Eq. 40* is more accurate than *Eq. 1*, then ascending the gradient for this model is likely to yield improved inference results. In practice, initializing with the inference from the fast filter assuming a linear model (e.g., *Eq. 30*) often resulted in nearly equally accurate inference, but inference assuming the above nonlinearity was far less robust than the inference assuming the linear model (not shown).

Using the fast filter to initialize the SMC filter. A sequential Monte Carlo (SMC) method to infer spike trains can incorporate this saturating nonlinearity, as well as other model extensions discussed earlier (Vogelstein et al. 2009). However, this SMC filter is not nearly as computationally efficient as the fast filter proposed here. Like the fast filter, the SMC filter estimates the model parameters in a completely unsupervised fashion, i.e., from the fluorescence observations, using an expectation-maximization algorithm (which requires iterating between computing the expected value of the hidden variables—$\boldsymbol{C}$ and $\boldsymbol{n}$—and updating the parameters). In Vogelstein and colleagues (2009), parameters for the SMC filter were initialized based on other data. Although effective, this initialization was often far from the final estimates and thus required a relatively large number of iterations (e.g., 20–25) before converging. Thus it seemed that the fast filter could be used to obtain an improvement to the initial parameter estimates, given an appropriate rescaling to account for the nonlinearity, thereby reducing the required number of iterations to convergence. Indeed, Fig. 8 shows how the SMC filter outperforms the fast filter on biological data and required only three to five iterations to converge on these data, given the initialization from the fast filter (which was typical). Note that the first few events of the spike train are individual spikes, resulting in relatively small fluorescence fluctuations, whereas the next events are actually spike doublets or triplets, causing a much larger fluorescence fluctuation. Only the SMC filter correctly infers the presence of isolated spikes in this trace, a frequently occurring result when the SNR is poor. Thus these two inference algorithms are complementary: the fast filter can be used for rapid, on-line inference, and for initializing the SMC filter, which can then be used to further refine the spike train

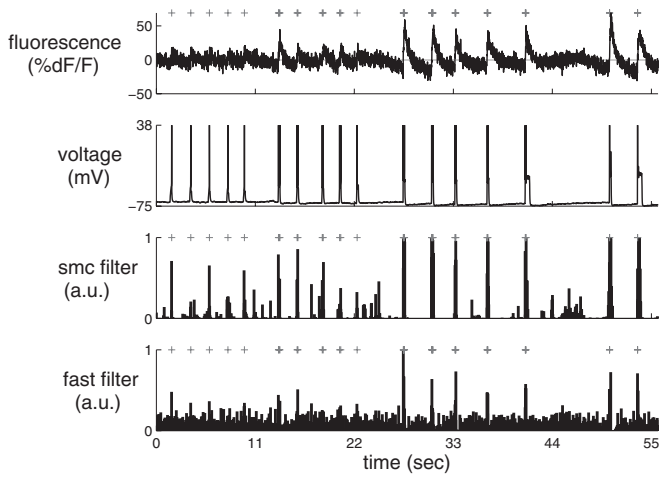

FIG. 8. In vitro data with SNR of only about 3 (estimated by dividing the fluorescent jump size by the SD of the baseline fluorescence) for single action potentials depicting the fast filter, effectively initializing the parameters for the sequential Monte Carlo (SMC) filter, significantly reducing the number of expectation-maximization iterations to convergence, using OGB-1. Note that whereas the fast filter clearly infers the spiking events in the end of the trace, those in the beginning of the trace are less clear. On the other hand, the SMC filter more clearly separates nonspiking activity from true spikes. Also note that the ordinate on the third panel corresponds to the inferred probability of a spike having occurred in each frame.

estimate. Importantly, although the SMC filter often outperforms the fast filter, the fast filter is more robust, meaning that it more often works "out of the box." This follows because the SMC filter operates on a highly nonlinear model that is not log-concave. Thus although the expectation-maximization algorithm used often converges to reasonable local maxima, it is not guaranteed to converge to global maxima and its performance in general will depend on the quality of the initial parameter estimates.

Spatial filter

In the preceding text, the filters operated on one-dimensional fluorescence traces. The raw data are in fact a time series of images that are first segmented into regions of interest (ROIs) and then (usually) spatially averaged to obtain a one-dimensional time series $\boldsymbol{F}$. In theory, one could improve the effective SNR of the fluorescence trace by scaling each pixel according to its SNR. In particular, pixels not containing any information about calcium fluctuations can be ignored and pixels that are partially anticorrelated with one another could have weights with opposing signs.

Figure 9 demonstrates the potential utility of this approach. The *top row* shows different depictions of an ROI containing a single neuron. On the *far left panel* is the true spatial filter for this neuron. This particular spatial filter was chosen based on experience analyzing both in vitro and in vivo movies; often, it seems that the pixels immediately around the soma are anticorrelated with those in the soma (MacLean et al. 2005; Watson et al. 2008). This effect is possibly due to the influx of calcium from the extracellular space immediately around the soma. The standard approach, given such a noisy movie, would be to first segment the movie to find an ROI corresponding to the soma of this cell and then spatially average all the pixels found to be within this ROI. The *second panel* shows this standard "boxcar spatial filter." The *third panel* shows the mean frame. The *fourth panel* shows the learned filter, using *Eq. 29* to estimate the spatial filter and background. Clearly, the learned filter is very similar to the mean filter and the true filter.

The *middle panels* of Figure 9 show the fluorescence traces obtained by background subtracting and then projecting each frame onto the corresponding spatial filter (black line) and true spike train (gray + symbols). The *bottom panels* show the inferred spike trains (black bars) using these various spatial filters and, again, the true spike train (gray + symbols). Although the

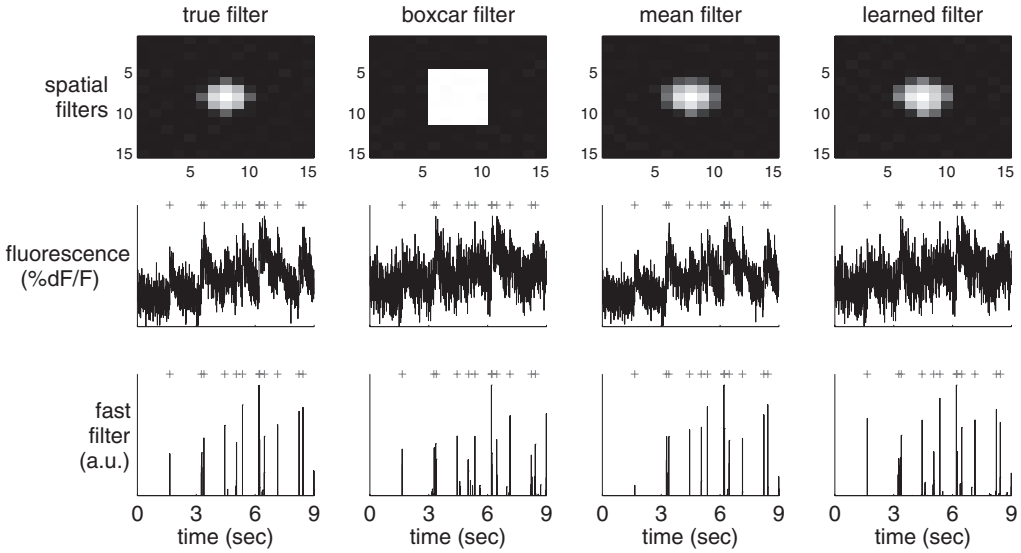

FIG. 9. A simulation demonstrating that using a better spatial filter can significantly enhance the effective SNR. The true spatial filter was a difference of Gaussians: a positively weighted Gaussian of small width and a negatively weighted Gaussian with larger width (both with the same center). Each column shows the spatial filter (*top*), one-dimensional fluorescence projection using that spatial filter (*middle*), and inferred spike train (*bottom*). From *left* to *right*, columns use the true, boxcar, mean, and learned spatial filter obtained using *Eq. 29*. Note that the learned filter's inferred spike train has fewer false positives and negatives than the boxcar and mean filters. Simulation parameters: $\tilde{\alpha} = \mathcal{N}(\mathbf{0}, 2\boldsymbol{I}) - 0.5\mathcal{N}(\mathbf{0}, 2.5\boldsymbol{I})$, where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates a 2-dimensional Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\tilde{\beta} = \mathbf{0}$, $\sigma = 0.2$, $\tau = 0.85$ s, $\lambda = 5$ Hz, $\Delta = 5$ ms, $T = 1{,}200$ time steps.

performance is very similar for all of them, the boxcar filter's inferred spike train is not as clean.

Overlapping spatial filters

The preceding text shows that if the ROI contains only a single neuron, the effective SNR can be enhanced by spatially filtering. However, this analysis assumes that only a single neuron is in the ROI. Often, ROIs are overlapping, or nearly overlapping, making the segmentation problem more difficult. Therefore it is desirable to have an ability to crudely segment, yielding only a few neurons in each ROI, and then spatially filter within each ROI to pick out the spike trains of each neuron. This may be achieved in a principled manner by generalizing the model as described in *Overlapping spatial filters* in METHODS. The true spatial filters of the neurons in the ROI are often unknown and thus must be estimated from the data. This problem may be considered a special case of blind source separation (Bell and Sejnowski 1995; Mukamel et al. 2009). Figure 10 shows that given reasonable assumptions of spiking correlations and SNR, multiple signals can be separated. Note that separation occurs even though the signal is significantly overlapping (*top panels*). To estimate the spatial filters, they are initialized using the boxcar filters (*middle panels*). After a few iterations, the spatial filters converge to very close approximation to the true spatial filters [compare true (*left*) and learned (*right*) spatial filters for the two neurons]. Note that both the true and learned spatial filters yield much improved spike inference relative to the boxcar filter. This suggests that even when spatial filters of multiple neurons are significantly overlapping, each spike train is potentially independently recoverable.

DISCUSSION

Summary

This work describes an algorithm that finds the approximate maximum a posteriori (MAP) spike train, given a calcium fluorescence movie. The approximation is required because finding the actual MAP estimate is not currently computationally tractable. Replacing the assumed Poisson distribution on spikes with an exponential distribution yields a log-concave optimization problem, which can be solved using standard gradient ascent techniques (such as Newton–Raphson). This exponential distribution has an advantage over a Gaussian distribution by restricting spikes to be positive, which improves inference quality (cf. Fig. 2), is a better approximation to a Poisson distribution with low rate, and imposes a sparse constraint on spiking. Furthermore, all the parameters can be estimated from only the fluorescence observations, obviating the need for joint electrophysiology and imaging (cf. Fig. 4). This approach is robust, in that it works "out of the box" on all the in vitro data analyzed (cf. Figs. 5 and 6). By using the special banded structure of the Hessian matrix of the log-posterior, this approximate MAP spike train can be inferred fast enough on standard computers to use it for on-line analyses of over 100 neurons simultaneously (cf. Fig. 7).

Finally, the fast filter is based on a biophysical model capturing key features of the data and may therefore be straightforwardly generalized in several ways to improve ac-

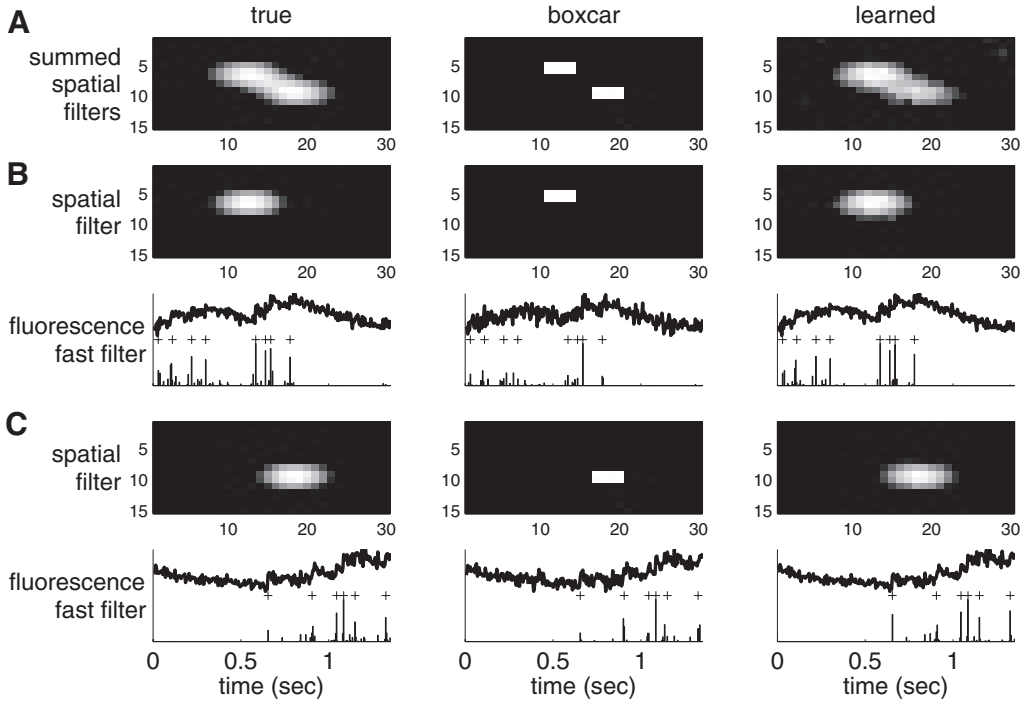

FIG. 10. Simulation showing that when 2 neurons' spatial filters are largely overlapping, learning the optimal spatial filters using *Eq. 36* can yield improved inference of the standard boxcar type filters. The 3 columns show the effect of the true (*left*), boxcar (*center*), and learned (*right*) spatial filters. *A*: the sum of the 2 spatial filters for each approach, clearly depicting overlap. *B*: the spatial filters (*top row*), one-dimensional fluorescence projection, and inferred spike train (*bottom row*) for one of the neurons. *C*: same as *B* for the other neuron. Note that the inferred spike trains when using the learned filter are close to optimal, unlike the boxcar filter. Simulation parameters: $\vec{\alpha}^1 = \mathcal{N}([-1, 0], 2\boldsymbol{I}) - 0.5\mathcal{N}([-1, 0], 2.5\boldsymbol{I})$, $\vec{\alpha}^2 = \mathcal{N}([1, 0], 2\boldsymbol{I}) - 0.5\mathcal{N}([1, 0], 2.5\boldsymbol{I})$, $\vec{\beta} = \mathbf{0}$, $\sigma = 0.02$, $\tau = 0.5$ s, $\lambda = 5$ Hz, $\Delta = 5$ ms, $T = 1{,}200$ time steps (not all time steps are shown).

curacy. Unfortunately, some of these generalizations do not improve inference accuracy, perhaps because of the exponential approximation. Instead, the fast filter output can be used to initialize the more general SMC filter (Vogelstein et al. 2009), to further improve inference quality (cf. Fig. 8). Another model generalization allows incorporation of spatial filtering of the raw movie into this approach (cf. Fig. 9). Even when multiple neurons are overlapping, spatial filters may be estimated to obtain improved spike inference results (cf. Fig. 10).

Alternate algorithms

This work describes but one specific approach to solving a problem that does not admit an exact solution that is computationally feasible. Several other approaches warrant consideration, including *1*) a Bayesian approach, *2*) a greedy approach, and *3*) different analytical approximations.

First, a Bayesian approach could use Markov Chain Monte Carlo methods to recursively sample spikes to estimate the full joint posterior distribution of the entire spike train, conditioned on the fluorescence data (Andrieu et al. 2001; Joucla et al. 2010; Mishchenko et al. 2010). Although enjoying several desirable statistical properties that are lacking in the current approach (such as consistency), the computational complexity of such an approach renders it inappropriate for the aims of this work.

Second, a common relatively expedient approximation to Bayesian sampling is a so-called greedy approach. Greedy algorithms are iterative, with each iteration adding another spike to the putative spike train. Each spike that is added is the most likely spike (thus the greedy term) or the one that most increases the likelihood of the fluorescence trace. Template matching, projection pursuit regression (Friedman and Stuetzle 1981), and matching pursuit (Mallat and Zhang 1993) are examples of such a greedy approach (the algorithm proposed by Grewe et al. (2010) could also be considered a special case of such a greedy approach).

Third, approximations other than the exponential distribution are possible. For instance, the Gaussian approximation is more appropriate for high firing rates, although in simulations, this more accurate approximation did not improve the Wiener filter output relative to the fast filter output (cf. Fig. 3). Perhaps the best approximation would use the closest log-concave relaxation to the Poisson model (Koenker and Mizera 2010). More formally, let $P(i)$ represent the Poisson mass at i and let $\ln Q$ be some concave density. Then, one could find the log-density Q such that Q maximizes $\sum_i P(i)Q(i) - \lambda \int \exp\{Q(x)\}\mathrm{d}x$ over the space of all concave Q. The first term corresponds to the log-likelihood, equivalent to the Kullback–Leibler divergence (Cover and Thomas 1991), and the second is a Lagrange multiplier to ensure that the density $\exp\{Q(x)\}$ integrates to unity. This is a convex problem because the space of all concave Q is convex and the objective function is concave in Q. In addition, it is easy to show that the optimal Q has to be piecewise linear; this means that one need not search over all possible densities, but rather, simply vary $Q(i)$ at the integers. Note that $\int \exp\{Q(x)\}\mathrm{d}x$ can be computed explicitly for any piecewise linear Q. This optimization problem can be solved using simple interior point methods and, in fact, the Hessian of the inner loop of the interior point method will be banded (because enforcing concavity of Q is a local constraint). This approximation could potentially be more accurate than our exponential approximation. Further, this approximation encourages integer solutions for n_t and is therefore of interest for future work.

The abovementioned three approaches may be thought of as complementary because each has unique advantages relative to the others. Both the greedy methods and the analytic approximations could potentially be used to initialize a Bayesian approach, possibly limiting the burn-in period, which can be computationally prohibitive in certain contexts. A greedy approach has the advantage of providing actual spike trains (i.e., binary sequences), unlike the analytic approximations. However, the actual spike trains could be quite far from the MAP spike train because greedy approaches, in general, have no guarantee of consistency. The analytic approximations, on the other hand, are guaranteed to converge to solutions close to the MAP spike train, where closeness is determined by the accuracy of the above approximation. Thus developing these distinct approaches and combining them is a potential avenue for further research.

Spatial filtering

Spatial filtering could be improved in a number of ways. For instance, pairing this approach with a crude but automatic segmentation tool to obtain ROIs would create a completely automatic algorithm that converts raw movies of populations of neurons into populations of spike trains. Furthermore, this filter could be coupled with more sophisticated algorithms to initialize the spatial filters when they are overlapping [for instance, principal component analysis (Horn and Johnson 1990) or independent component analysis (Mukamel et al. 2009)]. One could also use a more sophisticated model to estimate the spatial filters. One option would be to assume a simple parametric form of the spatial filter for each neuron (e.g., a basis set) and then merely estimate the parameters of that model. Alternately, one could regularize the spatial filters, using an elastic net type approach (Grosenick et al. 2009; Zou and Hastie 2005), to enforce both sparseness and smoothness.

Model generalizations

In this work, we made two simplifying assumptions that can easily be relaxed: *1*) instantaneous rise time of the fluorescence transient after a spike and *2*) constant background. In practice, often either or both of these assumptions are inaccurate. Specifically, genetic sensors tend to have a much slower rise time than that of organic dyes (Reiff et al. 2005). Further, the background often exhibits slow baseline drift due to movement, temperature fluctuations, laser power, and so forth, not to mention bleaching, which is ubiquitous for long imaging experiments. Both slow rise and baseline drift can be incorporated into our forward model using a straightforward generalization.

Consider the following illustrative example: the fluorescence rise time in a particular data set is quite slow, much slower than that of a single image frame. Thus fluorescence might be well modeled as the difference of two different calcium extrusion mechanisms, with different time constants. To model this scenario, one might proceed as follows: posit the existence of a two-dimensional time-varying signal, each like the calcium

APPENDIX A: PSEUDOCODE

Algorithm 1 Pseudocode for inferring the approximately most likely spike train, given fluorescence data. Note that the algorithm is robust to small variations ξ_z, ξ_n. The equations listed below refer to the most general equations in the text (simpler equations could be substituted when appropriate). Curly brackets, {·}, indicate comments.

initialize parameters, $\boldsymbol{\theta}$ (see *Initializing the parameters* in METHODS)
while convergence criteria not met **do**
  for $z = 1,0.1,0.01,\ldots,\xi_z$ do {interior point method to find $\hat{\boldsymbol{C}}$}
    Initialize $n_t = \xi_n$ for all $t = 1,\ldots,T$, $C_1 = 0$ and $C_t = \boldsymbol{\gamma} C_{t-1} + n_t$ for all $t = 2,\ldots,T$
    let $\boldsymbol{C}_z$ be the initialized calcium, and $\hat{P}_z$, be the posterior given this initialization
    while $\hat{P}_{z'} < \hat{P}_z$ **do** {Newton–Raphson with backtracking line searches}
      compute $\boldsymbol{g}$ using *Eq. 34*
      compute $\boldsymbol{H}$ using *Eq. 35*
      compute $\boldsymbol{d}$ using $\boldsymbol{H}\backslash\boldsymbol{g}$ {block-tridiagonal Gaussian elimination}
      let $\boldsymbol{C}_{z'} = \boldsymbol{C}_z + s\boldsymbol{d}$, where s is between 0 and 1, and $\hat{P}_{z'} > \hat{P}_z$ {backtracking line search}
    end while
  end for
  check convergence criteria
  update $\vec{\boldsymbol{\alpha}}$ and $\vec{\beta}$ using *Eq. 36* {only if spatial filtering}
  let σ be the root-mean square of the residual
  let $\lambda = T/(\Delta \Sigma_t \hat{n}_t)$
end while

signal assumed in the simpler models described earlier. Therefore each signal has a time constant and each signal is dependent on spiking. Finally, the fluorescence could be a weighted difference of the two signals. To formalize this model and to generalize it, let *1*) $\boldsymbol{X} = (X_1,\ldots,X_d)$ be a d-dimensional time-varying signal; *2*) $\boldsymbol{\Gamma}$ be a $d \times d$ dynamics matrix, where diagonal elements correspond to time constants of individual variables, and off-diagonal elements correspond to dependencies across variables; *3*) $\boldsymbol{A}$ be a d-dimensional binary column vector encoding whether each variable depends on spiking; and *4*) $\boldsymbol{\alpha}$ be a d-dimensional column vector of weights, determining the relative impact of each dimension on the total fluorescence signal. Given these conventions, we have the following generalized model:

$$F_t = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_t + \beta + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \tag{41}$$

$$\boldsymbol{X}_t = \boldsymbol{\Gamma} \boldsymbol{X}_{t-1} + \boldsymbol{A} n_t, \quad n_t \overset{\text{iid}}{\sim} \text{Poisson}(\lambda\Delta) \tag{42}$$

Note that this model simplifies to the model proposed earlier when $d = 1$. Because $\boldsymbol{X}$ is still Markov, all the theory developed above still applies directly for this model. There are, however, additional complexities with regard to identifiability. Specifically, the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{A}$ are closely related. Thus we enforce that $\boldsymbol{A}$ is a known binary vector, simply encoding whether a particular element responds to spiking. The matrix $\boldsymbol{\Gamma}$ will not be uniquely identifiable, for the same reason that γ was not identifiable, as described in *Learning the parameters* in METHODS. Thus we would assume $\boldsymbol{\Gamma}$ was known, a priori. Note that other approaches to dealing with baseline drift are also possible, such as letting β be a time-varying state: $\beta_t = \beta_{t-1} + \varepsilon_t$, where ε_t is a normal random variable with variance σ_β^2 that sets the effective drift rate. Both these models are the subjects of further development.

Concluding thoughts

In summary, the model and algorithm proposed in this work potentially provide a useful tool to aid in the analysis of calcium-dependent fluorescence imaging and establish the groundwork for significant further development.

APPENDIX B: WIENER FILTER

The Poisson distribution in *Eq. 4* can be replaced with a Gaussian instead of an exponential distribution, i.e., $n_t \overset{iid}{\sim} \mathcal{N}(\lambda\Delta, \lambda\Delta)$ that, when plugged into *Eq. 7*, yields:

$$\hat{\boldsymbol{n}} = \underset{n_t}{\text{argmax}} \sum_{t=1}^{T} \left[\frac{1}{2\sigma^2}(F_t - \alpha C_t - \beta)^2 + \frac{1}{2\lambda\Delta}(n_t - \lambda\Delta)^2 \right]. \tag{B1}$$

Note that since fluorescence integrates over Δ, it makes sense that the mean scales with Δ. Further, since the Gaussian here is approximating a Poisson with high rate (Sjulson and Miesenböck 2007), the variance should scale with the mean. Using the same tridiagonal trick as before, *Eq. 11*b can be solved using Newton–Raphson once (because this expression is quadratic in $\boldsymbol{n}$). Writing the above in matrix notation, substituting $C_t - \boldsymbol{\gamma} C_{t-1}$ for n_t, and letting $\alpha = 1$ yields:

$$\hat{\boldsymbol{C}} = \underset{\boldsymbol{C}}{\text{argmax}} -\frac{1}{2\sigma^2} \| \boldsymbol{F} - \boldsymbol{C} - \beta \boldsymbol{1}_T \|^2 - \frac{1}{2\lambda\Delta} \| \boldsymbol{M}\boldsymbol{C} - \lambda\Delta 1 \|^2, \tag{B2}$$

which is quadratic in $\boldsymbol{C}$. The gradient and Hessian are given by:

$$\boldsymbol{g} = -\frac{1}{\sigma^2}(\boldsymbol{C} - \boldsymbol{F} - \beta \boldsymbol{1}_T) - \frac{1}{\lambda\Delta}[(\boldsymbol{M}\hat{\boldsymbol{C}})^{\mathrm{T}} \boldsymbol{M} + \lambda\Delta \boldsymbol{M}^{\mathrm{T}} \boldsymbol{1}] \tag{B3}$$

$$\boldsymbol{H} = \frac{1}{\sigma^2} \boldsymbol{I} + \frac{1}{\lambda\Delta} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{M}. \tag{B4}$$

Note that this solution is the optimal linear solution, under the assumption that spikes follow a Gaussian distribution, and is often referred to as the Wiener filter, regression with a smoothing prior, or ridge regression (Boyd and Vandenberghe 2004). Estimating the parameters for this model follows a pattern similar to that described in *Learning the parameters* in METHODS.

ACKNOWLEDGMENTS

We thank V. Bonin for helpful discussions.

GRANTS

This work was supported by National Institute on Deafness and Other Communication Disorders Grant DC-00109 to J. T. Vogelstein; National Science Foundation (NSF) Faculty Early Career Development award, an Alfred P. Sloan Research Fellowship, and a McKnight Scholar Award to L.

Paninski; National Eye Institute Grant EY-11787 and the Kavli Institute for Brain Studies award to R. Yuste and the Yuste laboratory; and an NSF Collaborative Research in Computational Neuroscience award IIS-0904353, awarded jointly to L. Paninski and R. Yuste.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

REFERENCES

Andrieu C, Barat É, Doucet A. Bayesian deconvolution of noisy filtered point processes. *IEEE Trans Signal Process* 49: 134–146, 2001.

Bell AJ, Sejnowski TJ. An information-maximisation approach to blind separation and blind deconvolution. *Neural Comput* 7: 1129–1159, 1995.

Boyd S, Vandenberghe L. *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press, 2004.

Cover TM, Thomas JA. *Elements of Information Theory*: New York: Wiley–Interscience, 1991.

Cunningham JP, Shenoy KV, Sahani M. Fast Gaussian process methods for point process intensity estimation. In: *Proceedings of the 25th International Conference on Machine Learning* (ICML 2008). New York: IEEE Press, 2008, p. 192–199.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B Methodol* 39: 1–38, 1977.

Friedman JH, Stuetzle W. Projection pursuit regression. *J Am Stat Assoc* 76: 817–823, 1981.

Garaschuk O, Griesbeck O, Konnerth A. Troponin c-based biosensors: a new family of genetically encoded indicators for in vivo calcium imaging in the nervous system. *Cell Calcium* 42: 351–361, 2007.

Göbel W, Helmchen F. In vivo calcium imaging of neural network function. *Physiology* (Bethesda) 22: 358–365, 2007.

Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

Greenberg DS, Houweling AR, Kerr JND. Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nat Neurosci* 11: 749–751, 2008.

Grewe BF, Langer D, Kasper H, Kampa BM, Helmchen F. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. *Nat Methods* 7: 399–405, 2010.

Grosenick L, Anderson T, Smith SJ. Elastic source selection for in vivo imaging of neuronal ensembles. In: *Proceedings of the Sixth IEEE International Conference on Symposium on Biomedical Imaging: From Nano to Macro* (ISBI '09). New York: IEEE Press, 2009, p. 1263–1266.

Holekamp TF, Turaga D, Holy TE. Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy. *Neuron* 57: 661–672, 2008.

Horn R, Johnson C. *Matrix Analysis.* Cambridge, UK: Cambridge Univ. Press, 1990.

Huys QJM, Ahrens MB, Paninski L. Efficient estimation of detailed single-neuron models. *J Neurophysiol* 96: 872–890, 2006.

Ikegaya Y, Aaron G, Cossart R, Aronov D, Lampl I, Ferster D, Yuste R. Synfire chains and cortical songs: temporal modules of cortical activity. *Science* 304: 559–564, 2004.

Joucla S, Pippow A, Kloppenburg P, Pouzat C. Quantitative estimation of calcium dynamics from radiometric measurements: a direct, nonratioing method. *J Neurophysiol* 103: 1130–1144, 2010.

Kass R, Raftery A. Bayes factors. *J Am Stat Assoc* 90: 773–795, 1995.

Koenker R, Mizera I. Quasi-concave density estimation. *Ann Stat* 38: 2998–3027, 2010.

Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788–791, 1999.

Lin Y, Lee DD, Saul LK. Nonnegative deconvolution for time of arrival estimation. In: *Proceedings of the 2004 International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2004). New York: IEEE Press, 2004, p. 377–380.

Luo L, Callaway EM, Svoboda K. Genetic dissection of neural circuits. *Neuron* 57: 634–660, 2008.

MacLean J, Watson B, Aaron G, Yuste R. Internal dynamics determine the cortical response to thalamic stimulation. *Neuron* 48: 811–823, 2005.

Mallat S, Zhang Z. Matching pursuit with time-frequency dictionaries. *IEEE Trans Signal Process* 41: 3397–3415, 1993.

Mank M, Santos AF, Direnberger S, Mrsic-Flogel TD, Hofer SB, Stein V, Hendel T, Reiff DF, Levelt C, Borst A, Bonhoeffer T, Hbener M, Griesbeck O. A genetically encoded calcium indicator for chronic in vivo two-photon imaging. *Nat Methods* 5: 805–811, 2008.

Mao B, Hamzei-Sichani F, Aronov D, Froemke R, Yuste R. Dynamics of spontaneous activity in neocortical slices. *Neuron* 32: 883–898, 2001.

Markham J, Conchello J-A. Parametric blind deconvolution: a robust method for the simultaneous estimation of image and blur. *J Opt Soc Am A Opt Image Sci Vis* 16: 2377–2391, 1999.

Mishchenko Y, Vogelstein J, Paninski L. A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *Ann Appl Stat* http://www.imstat.org/aoas/nextissue.html.

Mukamel EA, Nimmerjahn A, Schnitzer MJ. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* 63: 747–760, 2009.

Nagayama S, Zeng S, Xiong W, Fletcher ML, Masurkar AV, Davis DJ, Pieribone VA, Chen WR. In vivo simultaneous tracing and Ca^{2+} imaging of local neuronal circuits. *Neuron* 53: 789–803, 2007.

O'Grady PD, Pearlmutter BA. Convolutive non-negative matrix factorisation with a sparseness constraint. In: *Proceedings of the International Workshop on Machine Learning for Signal Processing, 2006.* New York: IEEE Press, 2006, p. 427–432.

Paninski L, Ahmadian Y, Ferreira D, Koyama S, Rad KR, Vidne M, Vogenstein J, Wu W. A new look at state-space models for neural data. *J Comput Neurosci.* doi: 10.1007/s10827-009-0179-x, 1–20, 2009.

Pologruto TA, Yasuda R, Svoboda K. Monitoring neural activity and $[Ca^{2+}]$ with genetically encoded Ca^{2+} indicators. *J Neurosci* 24: 9572–9579, 2004.

Portugal LF, Judice JJ, Vicente LN. A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables. *Math Comput* 63: 625–643, 1994.

Press W, Teukolsky S, Vetterling W, Flannery B. *Numerical Recipes in C*. Cambridge, UK: Cambridge Univ. Press, 1992.

Reiff DF, Ihring A, Guerrero G, Isacoff EY, Joesch M, Nakai J, Borst A. In vivo performance of genetically encoded indictors of neural activity in flies. *J Neurosci* 25: 4766–4778, 2005.

Sasaki T, Takahashi N, Matsuki N, Ikegaya Y. Fast and accurate detection of action potentials from somatic calcium fluctuations. *J Neurophysiol* 100: 1668–1676, 2008.

Schwartz T, Rabinowitz D, Unni VK, Kumar VS, Smetters DK, Tsiola A, Yuste R. Networks of coactive neurons in developing layer 1. *Neuron* 20: 1271–1283, 1998.

Seeger M. Bayesian inference and optimal design for the sparse linear model. *J Machine Learn Res* 9: 759–813, 2008.

Sjulson L, Miesenböck G. Optical recording of action potentials and other discrete physiological events: a perspective from signal detection theory. *Physiology* (Bethesda) 22: 47–55, 2007.

Smetters D, Majewska A, Yuste R. Detecting action potentials in neuronal populations with calcium imaging. *Methods* 18: 215–221, 1999.

Vogelstein JT, Watson BO, Packer AM, Yuste R, Jedynak B, Paninski L. Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophys J* 97: 636–655, 2009.

Wallace DJ, zum Alten Borgloh SM, Astori S, Yang Y, Bausen M, Kgler S, Palmer AE, Tsien RY, Sprengel R, Kerr JND, Denk W, Hasan MT. Single-spike detection in vitro and in vivo with a genetic Ca^{2+} sensor. *Nat Methods* 5: 797–804, 2008.

Watson BO, MacLean JN, Yuste R. Up states protect ongoing cortical activity from thalamic inputs. *PLoS ONE* 3: e3971, 2008.

Wu MC-K, David SV, Gallant JL. Complete functional characterization of sensory neurons by system identification. *Annu Rev Neurosci* 29: 477–505, 2006.

Yaksi E, Friedrich RW. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca^{2+} imaging. *Nat Methods* 3: 377–383, 2006.

Yuste R, Konnerth A. Editors. *Imaging in Neuroscience and Development: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2005.

Yuste R, Katz LC. Control of postsynaptic Ca^{2+} influx in developing neocortex by excitatory and inhibitory neurotransmitters. *Neuron* 6: 333–344, 1991.

Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Statist Soc B Stat Methodol* 67: 301–320, 2005.

GENOMICS

The Predictive Capacity of Personal Genome Sequencing

Nicholas J. Roberts,[1]* Joshua T. Vogelstein,[2]* Giovanni Parmigiani,[3] Kenneth W. Kinzler,[1] Bert Vogelstein,[1†] Victor E. Velculescu[1†]

New DNA sequencing methods will soon make it possible to identify all germline variants in any individual at a reasonable cost. However, the ability of whole-genome sequencing to predict predisposition to common diseases in the general population is unknown. To estimate this predictive capacity, we use the concept of a "genometype." A specific genometype represents the genomes in the population conferring a specific level of genetic risk for a specified disease. Using this concept, we estimated the maximum capacity of whole-genome sequencing to identify individuals at clinically significant risk for 24 different diseases. Our estimates were derived from the analysis of large numbers of monozygotic twin pairs; twins of a pair share the same genometype and therefore identical genetic risk factors. Our analyses indicate that (i) for 23 of the 24 diseases, most of the individuals will receive negative test results; (ii) these negative test results will, in general, not be very informative, because the risk of developing 19 of the 24 diseases in those who test negative will still be, at minimum, 50 to 80% of that in the general population; and (iii) on the positive side, in the best-case scenario, more than 90% of tested individuals might be alerted to a clinically significant predisposition to at least one disease. These results have important implications for the valuation of genetic testing by industry, health insurance companies, public policy-makers, and consumers.

INTRODUCTION

As a result of continuing advances in high-throughput sequencing technologies (*1*–*4*), whole-genome sequencing will soon become an affordable approach to identify all sequence variants in an individual human. Recent evidence suggests that each human genome has more than 4.5 million sequence variants, some common, some infrequent (*5*). To date, several thousand genomic variants have been associated with human diseases, either as rare variants in Mendelian disorders or as common single-nucleotide polymorphisms in genome-wide association studies (GWAS) (*6*, *7*). Whole-genome or whole-exome sequencing has recently been used to identify new disease predisposing variants in various familial disorders, such as familial pancreatic cancer (*8*) and Miller syndrome (*9*). However, the potential utility of genome-wide sequencing for personalized medicine in the general population is unclear. Suppose, for example, that sequencing becomes sufficiently inexpensive that all individuals, at birth, could have their genomes sequenced at negligible cost. What fraction of the population would benefit from such sequencing? "Benefit" in this context is defined as receiving information indicating that the risk of disease is increased or decreased to a degree that would alter an individual's life-style or medical management.

On the surface, it might seem impossible to answer this question at present, because there are millions of genetic variants in every individual and the contribution of nearly all of these variants to any disease is unknown. However, there is one group of individuals in which this question can be immediately addressed: monozygotic twin pairs. If one twin of the pair has a disease, then the probability of the other twin developing that disease is dependent on the genome whenever that disease has some genetic component. We show below that when this logic is applied to a large numbers of twins, estimates of the maximal benefit of genome-wide sequencing in the general (non-twin) population can be made.

[1]Ludwig Center for Cancer Genetics and Therapeutics and The Howard Hughes Medical Institute at Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231, USA. [2]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA. [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA.
*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: vogelbe@jhmi.edu (B.V.); velculescu@jhmi.edu (V.E.V.)

RESULTS

Conceptual basis

The key to our analysis is the concept of a "genometype." We do not know the genomic sequences of the twin pairs analyzed in the studies described herein, but we do know that each twin pair shares a nearly identical genome (*10*) and that a genome confers a particular genetic risk to every disease. For each disease, we group genomes that confer identical genetic risks into genometypes. For example, genometypes could be grouped into 20 bins, with genometypes in bin 1 conferring zero genetic risk, genometypes in bin 2 conferring 3% genetic risk, genometypes in bin 3 conferring 10% genetic risk, etc. We can then estimate what distributions of genometypes in the population best reflect the observed monozygotic twin concordancy and discordancy for any given disease.

In twin studies on diseases, heritability (defined in Table 1) is generally based on the difference in the incidence of a disease in monozygotic versus dizygotic twins (*11*, *12*). Heritability reflects the average genetic contribution to disease in a twin population. We are interested in the distribution of genetic risks rather than the average. For example, a 30% average risk could reflect a small fraction of twin pairs with genometypes conferring high genetic risk or a larger fraction of twin pairs with genometypes conferring a moderate genetic risk. Among all the distributions of genometypes that are compatible with the twin epidemiologic data, we wished to find the distributions that maximized the potential clinical utility of identifying those genometypes by genomic sequencing.

Table 1. Definition of terms.

Genometype	A set of genomes that confer a specific genetic risk for a given disease
Genometype genetic risk (*r*)	The genetic risk conferred by a given genometype
Genometype frequency (*f*)	The frequency of a given genometype in the general population
Threshold	Minimum risk for a given disease considered to be clinically meaningful
Heritability (HER)	Proportion of phenotypic variance associated with genetic factors
Cohort risk (CR)	Risk of disease in the relevant twin cohort
Nongenetic risk (*e*)	Proportion of cohort risk due to nongenetic factors
Total risk	Sum of genetic risk conferred by a given genometype plus nongenetic risk
Relative risk	Ratio of total risk associated with a given genometype to cohort risk

Whole-genome sequencing-based tests, like any genetic test, can be informative in two ways: Negative and positive tests would indicate a substantially lower or higher risk, respectively, than that of the general population. The challenge is to define "substantially" in clinically meaningful and quantitative terms. An example might help put this challenge into perspective. Suppose a woman receives a whole-genome test result indicating that she has a 90% lifetime risk (the total risk over her entire life) of developing breast cancer. She may decide to have a prophylactic double mastectomy to prevent this outcome. Similarly, if the test indicated an 80% or even a 50% lifetime risk of developing breast cancer, she may consider mastectomy. On the other hand, if the test indicated only a 14% risk of developing breast cancer, then mastectomies would be considered by very few women, given that most women today do not opt for prophylactic mastectomies even though the lifetime risk of developing breast cancer in the general population is 12%.

This example illustrates that the risk threshold required for clinical utility represents a balance between the risk reduction afforded by an intervention and its negative consequences. A precedent exists for defining this threshold, in that the decision to implement genetic tests is often based on a positive predictive value (PPV) of at least 10%, implying that more than 1 in 10 patients with a positive test result are expected to develop disease (*13*). Although the choice of this threshold will depend on the specific intervention and should ideally be left to the individual, we use this 10% threshold for our population-level analyses of 20 of the 24 diseases analyzed (table S1). In the other four diseases (chronic fatigue syndrome, gastroesophageal reflux disorder, coronary heart disease–related death, and general dystocia), which occur at relatively high frequency in the population, this 10% threshold is inadequate to distinguish individuals with a substantially increased genetic risk from the rest of the population. For these four diseases (table S1), a more appropriate threshold corresponds to one conferring a genetic risk that is at least as great as that of the nongenetic component. Individuals with genometypes conferring this degree of genetic risk would therefore have a total risk at least twice as large as those without any genetic predisposing factors. This 2× threshold in relative risk is similar to those widely used as clinical benchmarks for common diseases (table S2) (*14*–*18*).

For whole-genome testing in healthy individuals, we thereby defined a threshold at which a positive test result would be clinically meaningful as follows. If the nongenetic risk was <5%, then the threshold was set at 10%. If the nongenetic risk was >5%, then the threshold was set at 2× the nongenetic contribution. Although we have used these particular thresholds in most of the examples described below, we also describe how these results varied when other thresholds were considered.

Twin data

We collated monozygotic twin pair data from the Swedish Twin Registry, Danish Twin Registry, Finnish Twin Cohort, Norwegian National Birth Registry, and the National Academy of Science–National Research World War II Veteran Twins Registry (*19*–*31*) (Table 2). From these registries, we selected data representing 24 diseases of diverse etiologies including autoimmune diseases, cancer, cardiovascular diseases, genitourinary diseases, neurological diseases, and obesity-associated diseases. Three of these conditions (coronary heart disease, cancer, and stroke) represent the leading causes of mortality in the United States, accounted for 54.2% of total deaths in 2007, and are therefore of major public health importance (*32*). The thresholds for a clinically meaningful test result, as defined above, were calculated from disease prevalence and nongenetic risks in the populations from which the twins were drawn (*19*–*31*) (Materials and Methods, Table 2, and table S2).

Mathematical model

We then developed computational methods to evaluate possible frequency (*f*) and genetic risk (*r*) combinations for a population containing 20 genometypes. Genometype frequency is defined as the proportion of twin pairs in the population that have a given genometype (Table 1). Genometype genetic risk is defined, for each disease, as the absolute increment in risk that an individual with that genometype will face compared to someone with no genetic risk at all (Table 1). For any combination of genometypes, each with a certain frequency and genetic risk, we obtain an expected distribution of disease-affected individuals among a monozygotic twin cohort. Many different combinations of genometype frequencies and genetic risks match the observed distributions in monozygotic twins; we are interested in those combinations (distributions) that maximize clinical utility, as noted above and further explained below. The mathematical framework for our study and associated statistical and technical issues are detailed in Materials and Methods.

Clinical implications

These analyses allowed us to address various measures of potential clinical utility. First, for each disease, what is the maximum and minimum fraction of patients with the disease that would receive a positive test, that is, a result indicating that they have a substantially increased or decreased risk, respectively, of that disease? The answers to this question are graphically shown in Fig. 1 for each of the 24 diseases (for three diseases, we present different answers for males and females, resulting in a total of 27 disease categories). As can be seen from Fig. 1, the maximum fraction of patients that would receive a positive test varies widely from disease to disease. Most of the patients (>50%) who would ultimately develop 13 of the 27 disease categories would not test positive, even in the best-case scenario. On the other hand, there were four disease categories—thyroid autoimmunity, type I diabetes, Alzheimer's disease, and coronary heart disease–related deaths in males—for which genetic tests might identify more than 75% of the patients who ultimately develop the disease. Genometype risk

Table 2. Population-based twin studies used for analysis. Disease prevalence in cohort [cohort risk (CR)] was determined as described in Materials and Methods. MZ, monozygotic.

Disease/condition	Sex	Number of MZ twin pairs	Number of MZ disease-concordant pairs	Number of MZ disease-discordant pairs	Disease prevalence in cohort (CR) (%)	Reference
Bladder cancer	Male and female	15,668	5	189	0.6	Lichtenstein *et al.* (*19*)
Breast cancer	Female	8,437	42	505	3.5	Lichtenstein *et al.* (*19*)
Colorectal cancer	Male and female	15,668	30	416	1.5	Lichtenstein *et al.* (*19*)
Leukemia	Male and female	15,668	2	103	0.3	Lichtenstein *et al.* (*19*)
Lung cancer	Male and female	15,668	18	296	1.1	Lichtenstein *et al.* (*19*)
Ovarian cancer	Female	8,437	3	125	0.8	Lichtenstein *et al.* (*19*)
Pancreatic cancer	Male and female	15,668	3	123	0.4	Lichtenstein *et al.* (*19*)
Prostate cancer	Male	7,231	40	299	2.6	Lichtenstein *et al.* (*19*)
Stomach cancer	Male and female	15,668	11	223	0.8	Lichtenstein *et al.* (*19*)
Thyroid autoimmunity	Male and female	284	7	17	5.5	Hansen *et al.* (*20*)
Type 1 diabetes	Male and female	4,307	3	20	0.3	Kaprio *et al.* (*21*)
Gallstone disease	Male and female	11,073	112	956	5.3	Katsika *et al.* (*22*)
Type 2 diabetes	Male and female	4,307	29	113	2.0	Kaprio *et al.* (*21*)
Alzheimer's disease	Male and female	398	2	8	1.5	Gatz *et al.* (*23*)
Dementia	Male and female	398	3	16	2.8	Gatz *et al.* (*23*)
Parkinson disease	Male	3,477	7	60	1.1	Tanner *et al.* (*24*)
Chronic fatigue	Female	1,803	133	526	22.0	Sullivan *et al.* (*25*)
	Male	1,426	48	266	12.7	Sullivan *et al.* (*25*)
Gastroesophageal reflux disorder	Female	1,260	63	284	16.3	Cameron *et al.* (*26*)
	Male	918	32	185	13.6	Cameron *et al.* (*26*)
Irritable bowel syndrome	Male and female	1,252	14	97	5.0	Bengtson *et al.* (*27*)
Coronary heart disease death	Female	2,004	97	424	15.4	Zdravkovic *et al.* (*28*)
	Male	1,640	153	451	23.1	Zdravkovic *et al.* (*28*)
Stroke-related death	Male and female	3,852	35	316	5.0	Bak *et al.* (*29*)
General dystocia	Female	928	40	173	13.6	Algovik *et al.* (*30*)
Pelvic organ prolapse	Female	3,376	34	157	3.3	Altman *et al.* (*31*)
Stress urinary incontinence	Female	3,376	13	87	1.7	Altman *et al.* (*31*)

and frequency distributions for all diseases are shown in table S3 and graphically for representative diseases in fig. S1.

We could also determine the maximum and minimum fraction of individuals in the population (rather than the fraction of patients with disease) who would receive positive test results for each disease. As shown in Fig. 2, this fraction is generally small, as expected, because the incidence of most diseases is relatively low. Are these negative tests, which would be received by the great majority of individuals for most diseases, informative? Negative tests could be informative to individual patients if they indicated a considerably lower total risk than would be assumed in the absence of testing. As can be seen from Fig. 3, though, negative tests are generally not very informative in the case of whole-genome sequencing, because such genetic tests are limited by the nongenetic component of risk. For 22 of the 27 disease categories studied, a negative test would not indicate a risk that is less than half that in the general population, even in the best-case scenario. This level of risk reduction is probably not sufficient to warrant changes of behavior, life-style, or preventative medical practices for these individuals (*33*, *34*). On the other hand, there was one disease category (Alzheimer's disease, Fig. 3) in which a negative test result might indicate as little as a ~12% relative risk of disease compared to the entire twin cohort, at least in the best-case scenario. Knowledge of such a reduced risk might be comforting and relieve anxiety, particularly to those with a family history of Alzheimer's disease.

What is the maximum fraction of individuals that could receive at least one positive test result, that is, a report indicating that he or she is at risk for at least 1 of the 24 diseases assessed? From the data depicted in Fig. 2, we estimate that >95% of men and >90% of women could receive at least one positive test result if the risk alleles were actually distributed in the way that produced maximal sensitivity in our model.

We assumed that the risk alleles for these 24 diseases were independent in these estimates; if they were not independent, then these figures represent overestimates. On the other hand, these frequencies may represent underestimates because there are a number of additional diseases with hereditary components that have not yet been studied in monozygotic twins or included in our analyses. At the very least, if we consider only distinct disease categories whose pathogenesis is unlikely to be shared, our analyses suggest that, in the best-case scenario, most of the tested individuals might be alerted to a clinically meaningful risk by whole-genome sequencing.

It was of interest to determine how the results described above varied with the threshold chosen for the analysis. For example, it might be argued that a threshold of 10% was too low for true clinical utility. Our analyses show that the maximum fraction of affected cases testing positive, as well as the maximum fraction of the total population that tests positive, is not changed much when the thresholds are changed to 20% (tables S4 and S5). With very high thresholds, however, both these measures of sensitivity decrease significantly (tables S4 and S5). Moreover, the maximum predictive value of a negative test drops precipitously at higher thresholds (table S6).

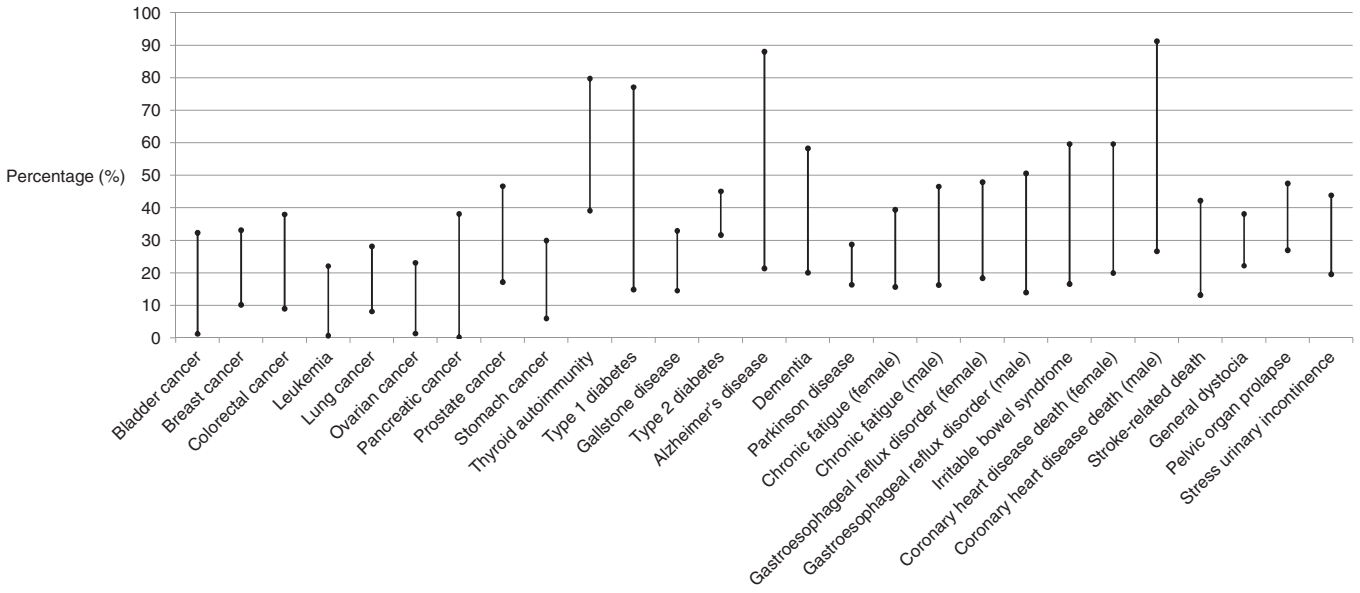

Fig. 1. The fraction of cases (that is, patients with disease) who would test positive by whole-genome sequencing. For each disease, the maximum and minimum fraction of cases that would test positive using the thresholds defined in table S1 are plotted.

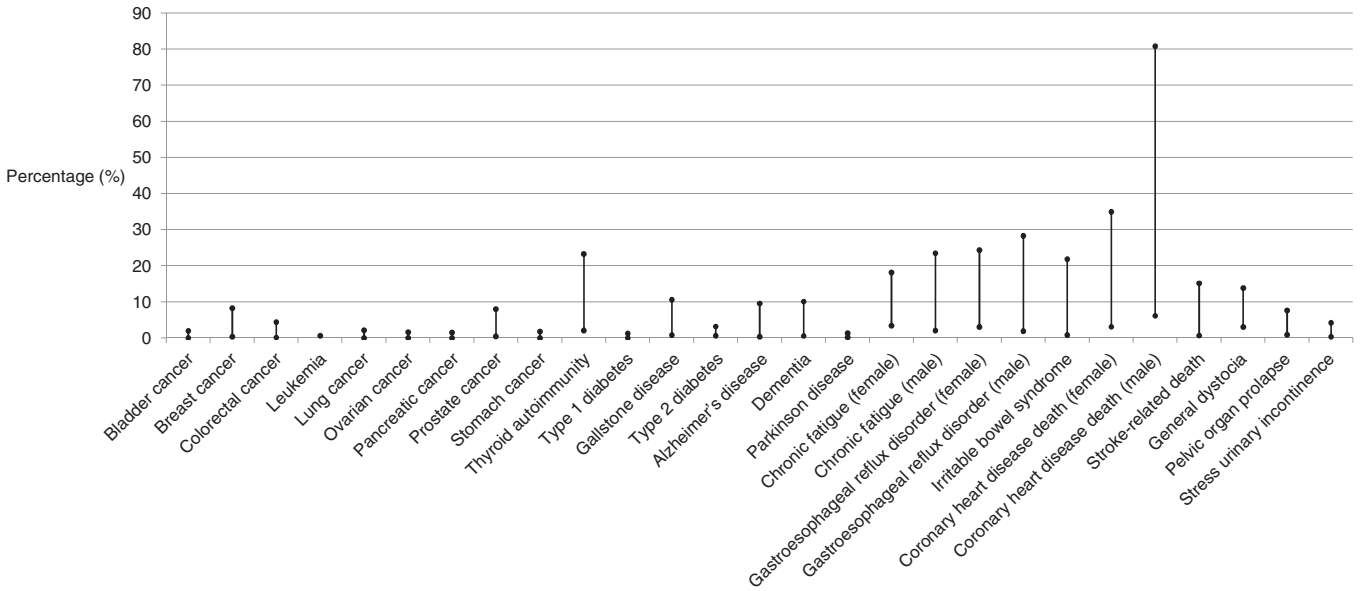

Fig. 2. Percentage of individuals in the general population who would test positive by whole-genome sequencing. For each disease, the maximum and minimum fraction of individuals in the population that would test positive using the thresholds defined in table S1 are plotted.

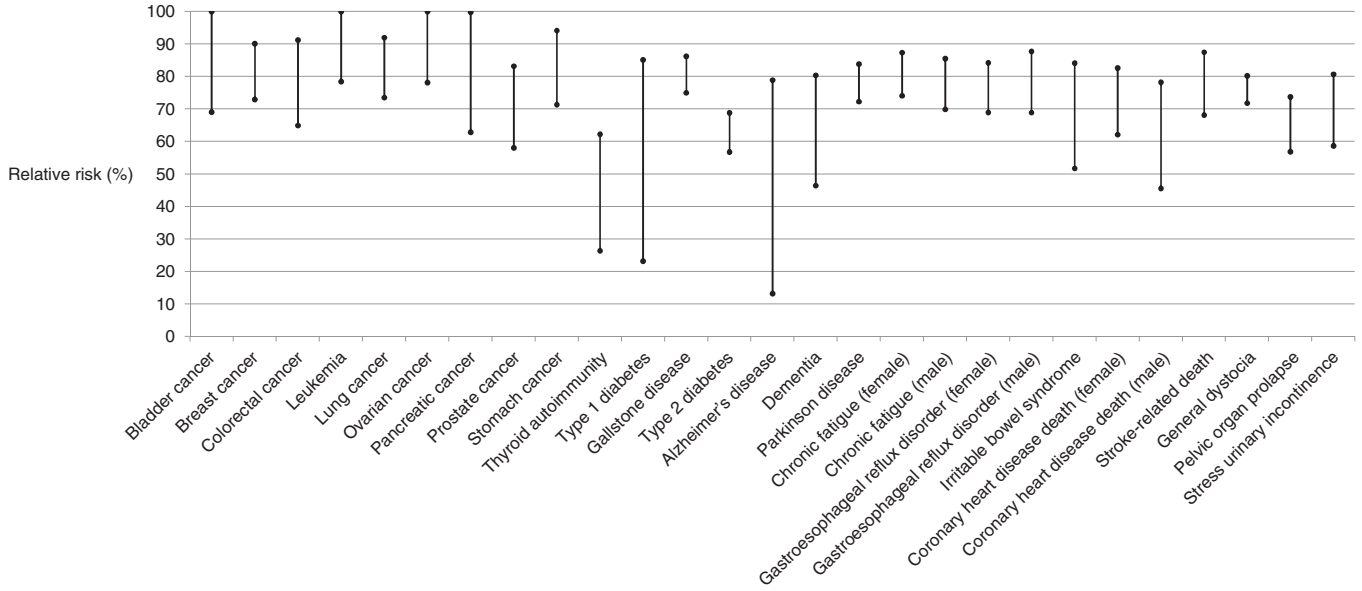

Fig. 3. Relative risk of disease in individuals testing negative by whole-genome sequencing. A relative risk of 100% represents the same risk as the general population, that is, the cohort risk. Relative risks were calculated using the genometype frequencies and genometype genetic risks that maximized or minimized sensitivity for disease detection.

DISCUSSION

The general public does not appear to be aware that, despite their very similar height and appearance, monozygotic twins in general do not always develop or die from the same maladies (*35*, *36*). This basic observation, that monozygotic twins of a pair are not always afflicted by the same maladies, combined with extensive epidemiologic studies of twins and statistical modeling, allows us to estimate upper and lower bounds of the predictive value of whole-genome sequencing.

On the negative side, our results show that most of the tested individuals would receive negative tests for most diseases (Fig. 2). Moreover, the predictive value of these negative tests would generally be small, because the total risk for acquiring the disease in an individual testing negative would be similar to that of the general population (Fig. 3). On the positive side, our results show that, at least in the best-case scenario, most of the patients might be alerted to a clinically meaningful risk for at least one disease through whole-genome sequencing.

These conclusions should be compared to other models as well as current knowledge about risk allele loci from GWAS (*5*–*7*, *37*–*39*, and references therein). In general, GWAS have shown that many loci can predispose to disease and that each risk allele confers a relatively small effect (*38*, *39*). For example, a recent analysis of large cohorts of individuals with colorectal cancer showed that only ~1.3% of phenotypic variance could be accounted for by the 10 loci discovered through GWAS (*40*). However, it could be argued that the relatively low level of utility that might be inferred from such studies is misleading. In particular, it is possible that a more complete knowledge of disease-associated variants and their epistatic relationships would be able to reliably predict who will and who will not develop disease in the general population. Modeling allows estimation of the maximum possible information that could be derived from such tests.

Several of our conclusions are based on the genometype frequency and risk distributions that would maximize the clinical utility of genetic testing, that is, are best-case scenarios. The actual frequency and risk distributions of genometypes in the population are not likely to be distributed in this way. Indeed, other distributions are also consistent with the monozygotic twin data on which our maxima are determined and all other distributions yield less clinical utility than those of the maxima. Moreover, in the real world, it is unlikely that the biomedical correlates of every genetic variant and the epistatic relationships among these variants will ever be completely known, or that the analytic validity of genetic testing will be perfect—as we assume in our ideal scenario.

Thus, our conclusions purposely overestimate the value of whole-genome sequencing that will be achieved—they represent an absolute upper bound that cannot be improved by improvements in technology or genetic knowledge. As a practical example of this principle, we estimate that a negative whole-genome sequencing–based test could indicate a nearly twofold decrease in risk for prostate cancer in men and a similar twofold decrease for urinary incontinence in women. But this twofold decrease would only apply in a world in which the risk alleles are distributed in a fashion that maximizes the sensitivity of whole-genome testing (Fig. 3). In the real world, the risk alleles are not likely to be distributed in this ideal fashion, and omniscience about every variant is not likely to be realized. Thus, the risk of these diseases in patients who test negative will likely be even more similar to that of the general population. For diseases with a lower heritable component, such as most forms of cancer, whole genome–based genetic tests will be even less informative. Thus, our results suggest that genetic testing, at its best, will not be the dominant determinant of patient care and will not be a substitute for preventative medicine strategies incorporating routine checkups and risk management based on the history, physical status, and life-style of the patient.

It is important to point out that our study focused on testing relatively common diseases in the general population and did not address the use of whole-genome sequencing to identify the genetic basis of rare monogenic diseases. In such unusual cases, it has already been

shown that whole-genome sequencing can prove highly informative [for example, (*8*, *9*)].

As with any model-based study, our conclusions have a number of caveats. Our analyses are based on data from twin studies and the assumptions made therein (*11*). Specifically, we do not model gene-environment interactions and rely on the prevalence of disease in the twin cohorts; this prevalence, as well as the operative nongenetic contributions, may differ from that in the general population. Although twins are likely to be representative of the general population, the estimates provided by our model could be improved through analyses of larger twin cohorts as these become available, as well as through a more complete phenotypic evaluation of twins of varying ethnicities. Another caveat is that our conclusions about potential utility are based on thresholds that represent a complex balance of personal choices, demographic influences, disease characteristics, and the clinical intervention(s) available. We have used a minimum 10% total risk and a minimum relative risk of 2 as the threshold in our analyses. Other thresholds may be more appropriate and meaningful for given situations, although the data in tables S4 to S6 show that our major conclusions are not altered much by the choice of threshold.

In sum, no result, including ours, can or should be used to conclude that whole-genome sequencing will be either useful or useless in an absolute sense. This utility will depend on the results of testing, the individual tested, and the perspectives of individuals and societies. What we hoped to accomplish with this study is to put the debate about the value of such sequencing in a mathematical and clinically relevant framework so that the potential merits and limitations of whole-genome sequencing, for any disease, can be quantitatively assessed. Recognition of these merits and limits can be useful to consumers, researchers, and industry, because they can minimize unrealistic expectations and foster fruitful investigations.

MATERIALS AND METHODS

Twin studies used for genotype analyses

We used data from twin studies arising from population-based twin registries to investigate the distribution of disease risk within the population (*19*–*31*). The registries in our study included the Swedish Twin Registry, Danish Twin Registry, Finnish Twin Cohort, Norwegian National Birth Registry, and the National Academy of Science–National Research Council World War II Veteran Twins Registry. Traits were chosen that represented diverse etiologies or were conditions of significant public health importance. We evaluated diseases in the following categories: autoimmune (type 1 diabetes and thyroid autoantibodies), neoplastic (breast, colorectal, and prostate cancer), cerebrovascular (coronary heart disease– and stroke-related death), genitourinary (general dystocia, pelvic organ prolapse, and urinary incontinence), unknown etiology (irritable bowel syndrome and chronic fatigue), neurological (Parkinson disease, Alzheimer's disease, and dementia), and obesity-associated (type 2 diabetes and gallstone disease).

To be included in our analyses, the following data had to be available for each twin study: (i) n_t—total number of monozygotic twin pairs where the disease status of each twin was known, (ii) n_c—number of disease-concordant monozygotic twin pairs, (iii) n_d—number of disease-discordant monozygotic twin pairs, (iv) n_h—number of healthy-concordant monozygotic pairs, and (v) heritability (HER)—calculated as the proportion of the polygenic liability variation associated with genetic factors.

Using the data from population-based twin studies, we define cohort risk (CR)—the fraction of people in the cohort that had the disease—as follows:

$$\mathrm{CR} = \frac{(2n_c + n_d)}{2n_t} \tag{1}$$

Model of the predictive capacity of personal genome sequencing

We define the following generative model that characterizes the joint distribution of an individual having a prespecified disease and a particular genometype. Each individual is characterized by (i) a binary (Bernoulli) random variable, Z, specifying whether he or she has the disease and (ii) a categorical random variable, G, indicating the genometype of the individual. This means that of the assumed extant genometypes, each individual can have only one of them. The joint distribution of both the disease and the genometype for an individual is given by $P(Z, G)$. This joint distribution decomposes into a product of the likelihood of getting the disease given the genometype, $P(Z|G)$, and the prior probability of having the genometype, $P(G)$:

$$P(Z,G) = P(Z|G)P(G) \tag{2}$$

Thus, to proceed, we specify both the likelihood function, $P(Z|G)$, and the prior, $P(G)$. As mentioned above, G is a categorical random variable taking values $g_1, g_2, \ldots, g_d$, each of which with some probability. Therefore, we have

$$P(G = g_i) = f_i \tag{3}$$

for all $i = 1, 2, \ldots, d$. In words, a person can have one of the d assumed extant genometypes, and the probability of having genometype i is given by f_i.

The probability of having the disease given a genometype is $q_i = e + r_i$. Thus, q_i is the sum of a nongenetic risk, e, that is assumed to be constant for the whole population, and genetic risk, r_i (note that $0 \le q_i \le 1$). Nongenetic risk (e) is the proportion of people in the population that would get the disease if all had the most favorable genometype possible. Nongenetic risk includes all factors that are not inherited, including environmental exposures (for example, diet and carcinogens), epigenetic alterations, and stochastic influences. We estimated it as follows: $e = \mathrm{CR}(1 - \mathrm{HER})$ (see below). This model assumes that all risks are either nongenetic or genetic, that is, no interactions. We require that the unknown parameters, r_i, must be between 0 and $1 - e$, for all i. Therefore, for a given genometype, the likelihood term for genometype i is given by

$$P(Z = z|G = g_i) = \begin{cases} e + r_i, & \text{if } z = 1 \\ 1 - e - r_i, & \text{if } z = 0 \end{cases} \tag{4}$$

Thus, the joint distribution of disease and genometype can be written as follows:

$$P(Z = z, G = g_i) = f_i(e + r_i)^z(1 - e - r_i)^{1-z},\ z \in \{0, 1\},\ g_i \in \{g_1, \ldots, g_d\} \tag{5}$$

If the available data included the genometype and disease status of each individual, then inferring estimates of the parameters,

$r = (r_1, \ldots, r_d)$, and $f = (f_1, \ldots, f_d)$, would be relatively straightforward. However, the available data include only the disease status of monozygotic twins. When considering monozygotic twins, these represent observations of disease status in two individuals with identical genometypes. Therefore, we can describe a joint distribution for monozygotic twins having a disease or not. Let $Z_j = Z(X_j)$ be the Bernoulli random variable indicating whether a particular individual has disease, and let $Z_k = Z(X_k)$ be the Bernoulli random variable for the co-twin. Similarly, let $G_j = G(X_j)$ and $G_k = G(X_k)$ be categorical random variables indicating whether twin j or k of a pair has some particular genometype. The distribution of disease within monozygotic twins can be divided into three distinct groups, namely, disease-concordant, disease-discordant, and healthy-concordant pairs.

The probability of disease-concordant monozygotic twins is given by

$$\begin{aligned} P(Z_j = Z_k = 1 | G_j = G_k) &= \sum_i P(Z_j = Z_k = 1 | G_j = G_k = g_i) P(G_j = G_k = g_i) && (6a) \\ &= \sum_i P(Z_j = 1 | G_j = g_i) P(Z_k = 1 | G_k = g_i) P(G_j = G_k = g_i) && (6b) \\ &= \sum_i (e + r_i)^2 f_i && (6c) \end{aligned}$$

Similarly, the probability of healthy-concordant monozygotic twin pairs is given by

$$\begin{aligned} P(Z_j = Z_k = 0 | G_j = G_k) &= \sum_i P(Z_j = Z_k = 0 | G_j = G_k = g_i) P(G_j = G_k = g_i) && (7a) \\ &= \sum_i (1 - e - r_i)^2 f_i && (7b) \end{aligned}$$

The probability of monozygotic twin pairs discordant for disease is given by

$$P(Z_j \neq Z_k | G_j = G_k) = 2 \sum_i (e + r_i)(1 - e - r_i) f_i \quad (8)$$

Optimization

For each disease, let n_c, n_h, and n_d correspond to the number of concordant diseased, healthy, and discordant twin pairs, respectively. Assuming that there are d genometypes, the expected number of twin pairs of each of the three types is simply the total number of twin pairs times the probability of being each kind of twin pair:

$$\mathrm{E}[n_c] = n_t \sum_{i \in [d]} (e + r_i)^2 f_i \quad (9)$$

$$\mathrm{E}[n_c] = n_t \sum_{i \in [d]} (e + r_i)^2 f_i \quad (10)$$

$$\mathrm{E}[n_d] = n_t \sum_{i \in [d]} 2(e + r_i)(1 - e - r_i) f_i \quad (11)$$

Because we are interested in the limits of utility of genetic testing, we search for a parameter set that maximizes or minimizes the fraction of patients who will receive a positive test result, given certain constraints. Formally, we define the positive fraction (PF) as the proportion of cases that have a genometype sufficient to change clinical action. In our notation:

$$\mathrm{PF}(t, e; f, r) = \frac{\sum_{i \in [d] | r_i > t} f_i [(e + r_i)^2 + (e + r_i)(1 - e - r_i)]}{\sum_{i \in [d]} f_i [(e + r_i)^2 + (e + r_i)(1 - e - r_i)]} \quad (12)$$

where t is the genetic risk required for a person to be at the threshold required for clinical utility and d is the maximum number of genometypes under consideration. The thresholds for each disease are provided in table S1, and for each disease, t is defined as this threshold minus e.

We therefore seek to solve the following optimization problem, for each disease:

$$\underset{f,\, r}{\text{maximize}} \quad \mathrm{PF}(t, e; f, r) \quad (13)$$

$$\text{subject to } f_i \geq 0, \sum_i f_i = 1, r_i \in (0, 1 - e), \sum_{x \in \{c,h,d\}} (\hat{n}_x - \mathrm{E}[n_x])^2 \leq 0.25 \quad (14)$$

where Eq. 14 enforces that none of the residual errors can be larger than 0.5. The estimated number of twin pairs, n_x, is the estimated number of twin pairs of each type obtained by plugging the estimated parameters into Eqs. 9 to 11. This is therefore a quadratically constrained optimization problem. We use the following algorithm to obtain a local optimum.

For $d' = 2$, that is, starting with $d' = 2$ genometypes, we implement a grid search over the parameter space and select the parameters that maximize the likelihood over a constrained search space. Let $\theta = (f, r)$ and Θ be the set of all θs under consideration, as defined by the feasible region specified in Eq. 14. We then discretize this space into 9 bins for each element of f and 100 bins for each element of r and denote $P(Z|G)$ by $P_\theta(Z|G)$ to emphasize the dependence of the joint distribution on the parameter. Thus, we aim to solve the following optimization problem:

$$\hat{\theta}^{(2)} = \underset{\theta \in \Theta}{\arg\max} \prod_{i,j} P_\theta(Z_j, Z_k | G_j = G_k) \quad (15)$$

where $\hat{\theta}^{(d')} = (\hat{f}^{(d')}, \hat{r}^{(d')})$ is the parameter estimate assuming only d' genometypes. For each $d' = 3, \ldots, 20$, we seek to solve the above optimization problem. To initialize, we pad the previous solution with zeros, yielding $\hat{f}^{(d'+1)}_{(0)} = (\hat{f}^{(d')}, 0)$ and similarly for $\hat{r}^{(d'+1)}_{(0)}$. Then, we use MATLAB's fmincon to find a local maximum of PF given the constraints. If no improvement in PF is obtained for $d' + 1$ genometypes using the default "padded" initialization, we try randomly initializing. We stop trying random initializations if any of the following criteria are met: (i) if we find an improvement in PF with the constraints satisfied, (ii) if we reach 100% PF, or (iii) if we reach 15 random initializations. If criterion (i) is met, we denote the parameters achieving the improvement $\hat{\theta}^{(d'+1)}$ and then increment d' and continue. If criterion (ii) is met, we stop incrementing d', because we have achieved the maximum possible PF, so adding additional genometypes cannot possibly

maximize it further. If criterion (iii) is met, we let $\hat{\theta}^{(d'+1)} = \hat{\theta}^{(d'+1)}_{(0)}$; that is, we let our final estimate for $d' + 1$ simply be our estimate for d' padded with a zero. We then increment d'.

We repeat the above approach for each disease. The parameters that we determined using this approach to maximize PF were then used to estimate the percentage of the population testing positive for a given disease, as well as the relative risk of disease for those individuals testing negative, as defined below. We apply this approach separately for each disease, thus assuming independence. To find the minimum PFs compatible with the twin data, we used a similar procedure.

Relative risk of disease if testing negative

We determined the relative risk of disease of individuals whose whole-genome sequencing tests were negative after maximizing or minimizing the sensitivity (PF) of the test. Disease risk in the population testing negative (DR_{neg}) is the ratio of the number of disease cases testing negative to the number of individuals in the population testing negative:

$$DR_{neg} = \frac{(2n_c + n_d)(1 - PF)}{2n_t \sum_{i \in [d] r_i < t} f_i} \quad (16)$$

To determine the relative risk of disease if testing negative (RR_{neg}), we calculated the ratio of disease risk of individuals testing negative to the disease risk in the twin cohort (CR):

$$RR_{neg} = \frac{DR_{neg}}{CR} \quad (17)$$

Calculation of relative risks

We defined relative risk in table S1 as the minimum total risk of individuals with genometypes carrying a given genetic risk compared to the total risk of individuals with genometypes carrying a genetic risk of 0% (that is, determined solely by nongenetic factors). The minimum total risk was determined using the standard 10% risk threshold described in the text as well as others (tables S4 to S6). In all cases,

$$RR = \frac{PPV + [CR(1 - HER)]}{CR(1 - HER)} \quad (18)$$

Other parameters and models

Equation 14 enforces that none of the residual errors can be larger than 0.5, such that upon rounding we obtain a perfect fit. Changing this parameter from 0.5 to 0.01 did not alter the PFs depicted in Fig. 1 for any disease.

Instead of maximizing PFs, we also determined the distributions of genometype risks (r_i) and frequencies (f_i) that would minimize the relative risk of disease of individuals whose whole-genome sequencing tests were negative. This independent optimization yielded results nearly identical to those reported in Figs. 1 to 3.

As noted above, we estimated the nongenetic risk as $e = CR(1 - HER)$. This risk is somewhat higher than that derived from the standard liability threshold (LT) model. However, it has recently been shown that the LT model underestimates the nongenetic contribution to disease because it does not take into account synergistic interactions among genes (*41*). The model described herein does not make any assumptions about the nature of the interactions between genes, such as additivity. However, the LT model can also be used to approximate the maximum capacity of whole-genome sequencing to detect individuals at predefined risks under certain simplifying assumptions about the distribution of risk alleles in the population. The PF predictions from the LT model using 10% thresholds are provided in table S4 and can be compared to the results of the current model with 10% thresholds (table S4).

Finally, our model can be used to calculate the potential clinical utility of whole-genome sequencing under any assumption about the proportion of nongenetic contributions to disease risk, or estimates thereof. Representative values for each disease, with nongenetic contributions ranging from 10 to 90%, are provided in table S7.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/4/133/133ra58/DC1

Fig. S1. Graphical representation of genotype frequency and risk distributions for (A) leukemia, (B) Alzheimer's disease, and (C) pancreatic cancer.

Table S1. Examples of known risk factors for common human diseases.

Table S2. Thresholds and other parameters used to analyze each disease.

Table S3. The risks and frequencies of each of the 20 genometypes providing maximum sensitivity (PF) for detection of each disease.

Table S4. Percentage of cases (that is, individuals with disease) testing positive with whole-genome sequencing at varying risk thresholds or with the liability threshold (LT) model.

Table S5. Percentage of population testing positive with whole-genome sequencing at varying risk thresholds.

Table S6. Relative risk of disease if testing negative with whole-genome sequencing at varying risk thresholds.

Table S7. Percentage of cases testing positive with whole-genome sequencing at varying estimates of nongenetic contributions.

REFERENCES AND NOTES

1. D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, A. J. Smith, Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

2. R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig,

C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun, J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church, C. A. Reid, Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).

3. J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, S. Turner, Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
4. M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, J. M. Rothberg, Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
5. K. A. Frazer, S. S. Murray, N. J. Schork, E. J. Topol, Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
6. E. T. Cirulli, D. B. Goldstein, Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
7. T. A. Manolio, Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 (2010).
8. S. Jones, R. H. Hruban, M. Kamiyama, M. Borges, X. Zhang, D. W. Parsons, J. C. Lin, E. Palmisano, K. Brune, E. M. Jaffee, C. A. Iacobuzio-Donahue, A. Maitra, G. Parmigiani, S. E. Kern, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, J. R. Eshleman, M. Goggins, A. P. Klein, Exomic sequencing identifies *PALB2* as a pancreatic cancer susceptibility gene. *Science* **324**, 217 (2009).
9. S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, M. J. Bamshad, Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
10. C. E. Bruder, A. Piotrowski, A. A. Gijsbers, R. Andersson, S. Erickson, T. Diaz de Ståhl, U. Menzel, J. Sandgren, D. von Tell, A. Poplawski, M. Crowley, C. Crasto, E. C. Partridge, H. Tiwari, D. B. Allison, J. Komorowski, G. J. van Ommen, D. I. Boomsma, N. L. Pedersen, J. T. den Dunnen, K. Wirdefeldt, J. P. Dumanski, Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* **82**, 763–771 (2008).
11. F. V. Rijsdijk, P. C. Sham, Analytic approaches to twin data using structural equation models. *Brief. Bioinform.* **3**, 119–133 (2002).
12. P. M. Visscher, W. G. Hill, N. R. Wray, Heritability in the genomics era—Concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
13. D. L. Clarke-Pearson, Clinical practice. Screening for ovarian cancer. *N. Engl. J. Med.* **361**, 170–177 (2009).
14. P. G. Kopelman, Obesity as a medical problem. *Nature* **404**, 635–643 (2000).
15. W. C. Willett, W. H. Dietz, G. A. Colditz, Guidelines for healthy weight. *N. Engl. J. Med.* **341**, 427–434 (1999).
16. A. J. Alberg, J. G. Ford, J. M. Samet; American College of Chest Physicians, Epidemiology of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest* **132**, 29S–55S (2007).
17. A. Ott, A. J. Slooter, A. Hofman, F. van Harskamp, J. C. Witteman, C. Van Broeckhoven, C. M. van Duijn, M. M. Breteler, Smoking and risk of dementia and Alzheimer's disease in a population-based cohort study: The Rotterdam Study. *Lancet* **351**, 1840–1843 (1998).
18. J. He, L. G. Ogden, L. A. Bazzano, S. Vupputuri, C. Loria, P. K. Whelton, Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study. *Arch. Intern. Med.* **161**, 996–1002 (2001).
19. P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, K. Hemminki, Environmental and heritable factors in the causation of cancer—Analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
20. P. S. Hansen, T. H. Brix, I. Iachine, K. O. Kyvik, L. Hegedüs, The relative importance of genetic and environmental effects for the early stages of thyroid autoimmunity: A study of healthy Danish twins. *Eur. J. Endocrinol.* **154**, 29–38 (2006).
21. J. Kaprio, J. Tuomilehto, M. Koskenvuo, K. Romanov, A. Reunanen, J. Eriksson, J. Stengård, Y. A. Kesäniemi, Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* **35**, 1060–1067 (1992).
22. D. Katsika, A. Grjibovski, C. Einarsson, F. Lammert, P. Lichtenstein, H. U. Marschall, Genetic and environmental influences on symptomatic gallstone disease: A Swedish study of 43,141 twin pairs. *Hepatology* **41**, 1138–1143 (2005).
23. M. Gatz, N. L. Pedersen, S. Berg, B. Johansson, K. Johansson, J. A. Mortimer, S. F. Posner, M. Viitanen, B. Winblad, A. Ahlbom, Heritability for Alzheimer's disease: The study of dementia in Swedish twins. *J. Gerontol. A Biol. Sci. Med. Sci.* **52**, M117–M125 (1997).
24. C. M. Tanner, R. Ottman, S. M. Goldman, J. Ellenberg, P. Chan, R. Mayeux, J. W. Langston, Parkinson disease in twins: An etiologic study. *JAMA* **281**, 341–346 (1999).
25. P. F. Sullivan, B. Evengård, A. Jacks, N. L. Pedersen, Twin analyses of chronic fatigue in a Swedish national sample. *Psychol. Med.* **35**, 1327–1336 (2005).
26. A. J. Cameron, J. Lagergren, C. Henriksson, O. Nyren, G. R. Locke III, N. L. Pedersen, Gastroesophageal reflux disease in monozygotic and dizygotic twins. *Gastroenterology* **122**, 55–59 (2002).
27. M. B. Bengtson, T. Ronning, M. H. Vatn, J. R. Harris, Irritable bowel syndrome in twins: Genes and environment. *Gut* **55**, 1754–1759 (2006).
28. S. Zdravkovic, A. Wienke, N. L. Pedersen, M. E. Marenberg, A. I. Yashin, U. De Faire, Heritability of death from coronary heart disease: A 36-year follow-up of 20 966 Swedish twins. *J. Intern. Med.* **252**, 247–254 (2002).
29. S. Bak, D. Gaist, S. H. Sindrup, A. Skytthe, K. Christensen, Genetic liability in stroke: A long-term follow-up study of Danish twins. *Stroke* **33**, 769–774 (2002).
30. M. Algovik, E. Nilsson, S. Cnattingius, P. Lichtenstein, A. Nordenskjöld, M. Westgren, Genetic influence on dystocia. *Acta Obstet. Gynecol. Scand.* **83**, 832–837 (2004).
31. D. Altman, M. Forsman, C. Falconer, P. Lichtenstein, Genetic influence on stress urinary incontinence and pelvic organ prolapse. *Eur. Urol.* **54**, 918–922 (2008).
32. J. Xu, K. D. Kochanek, S. L. Murphy, B. Tejada-Vera, Deaths: Final data for 2007. *Natl. Vital. Stat. Rep.* **58**, 1–135 (2010).
33. J. Audrain, N. R. Boyd, J. Roth, D. Main, N. F. Caporaso, C. Lerman, Genetic susceptibility testing in smoking-cessation treatment: One-year outcomes of a randomized trial. *Addict. Behav.* **22**, 741–751 (1997).
34. E. Sabaté, *Adherence to Long-Term Therapies: Evidence for Action* (World Health Organization, Geneva, 2003).
35. A. H. Wong, I. I. Gottesman, A. Petronis, Phenotypic differences in genetically identical organisms: The epigenetic perspective. *Hum. Mol. Genet.* **14 Spec. No. 1**, R11–R18 (2005).
36. "Identical twins not as identical as believed," *ScienceDaily*, 15 February 2008; http://www.sciencedaily.com/releases/2008/02/080215121214.htm.
37. N. R. Wray, M. E. Goddard, P. M. Visscher, Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
38. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, T. A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009).
39. N. R. Wray, J. Yang, M. E. Goddard, P. M. Visscher, The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6**, e1000864 (2010).
40. A. Tenesa, M. G. Dunlop, New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat. Rev. Genet.* **10**, 353–358 (2009).
41. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1193–1198 (2012).

Acknowledgments: We thank N. Wray and D. Geman for critical comments regarding the manuscript, and K. Kinzler for technical assistance. **Funding:** The project was supported by NIH grant CA121113, the Virginia & D. K. Ludwig Fund for Cancer Research, and American Association for Cancer Research Stand Up To Cancer–Dream Team Translational Cancer Research Grant. **Author contributions:** N.J.R., J.T.V., G.P., K.W.K., B.V., and V.E.V. designed the study; N.J.R., J.T.V., and V.E.V. generated and analyzed the data; N.J.R., J.T.V., B.V., and V.E.V. wrote the manuscript. **Competing interests:** B.V., K.W.K., and V.E.V. are co-founders of Inostics and Personal Genome Diagnostics and are members of their Scientific Advisory Boards. K.W.K., B.V., and V.E.V. own Inostics and Personal Genome Diagnostics stock, which is subject to certain restrictions under University policy. The terms of these arrangements are managed by the Johns Hopkins University in accordance with its conflict-of-interest policies. G.P. is on the scientific advisory board of Counsyl.

Submitted 25 October 2011
Accepted 23 March 2012
Published 9 May 2012
10.1126/scitranslmed.3003380

Citation: N. J. Roberts, J. T. Vogelstein, G. Parmigiani, K. W. Kinzler, B. Vogelstein, V. E. Velculescu, The predictive capacity of personal genome sequencing. *Sci. Transl. Med.* **4**, 133ra58 (2012).

Science Translational Medicine

The Predictive Capacity of Personal Genome Sequencing

Nicholas J. Roberts, Joshua T. Vogelstein, Giovanni Parmigiani, Kenneth W. Kinzler, Bert Vogelstein and Victor E. Velculescu

Sci Transl Med **4**, 133ra58133ra58.
First published 2 April 2012
DOI: 10.1126/scitranslmed.3003380

Is It All in Your Genes?

Imagine that everyone at birth could have their whole genome sequenced at negligible cost. Surely, this must be a worthwhile endeavor, given the list of luminaries that have already had this sequencing completed. But how well will such tests perform? Will we be able to predict what diseases individuals will develop, and die from, right from birth?

In a study that seeks to answer these questions, Vogelstein and his colleagues present an unbiased assessment of the capacity of whole-genome sequencing to provide clinically relevant information assuming that future research will allow us to understand the significance of every genetic variant. Using previously published data on twins and a new mathematical framework, Vogelstein and his co-workers were able to estimate the maximum capacity of whole-genome sequencing to predict the risk for 24 relatively common diseases. They show that most of the tested individuals could be alerted to a predisposition to at least one disease. However, in any given individual, whole-genome sequencing will be relatively uninformative for most diseases, because the estimated risk of developing these diseases will be similar to that of the general population. Thus, for most patients, genetic testing will not be the dominant determinant of patient care and will not be a substitute for preventative medicine strategies incorporating routine checkups and risk management based on the history, physical status, and life-style of the individual.

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Bioinformatics, Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan. [7]Comparative Genomics Laboratory, National Institute of Genetics, 411-8540 Yata 1111, Mishima, Shizuoka, 411-8540, Japan. [8]Computational Biology Group, IIDMM, University of Cape Town Faculty of Health Sciences, Cape Town, 7925, South Africa. [9]Department of Biochemistry and Biophysics, Texas A&M University, 328B TAMU, College Station, TX 77843, USA. [10]Department of Biochemistry and Molecular Biology, Egerton University, Post Office Box 536, Njoro, Kenya. [11]Department of Biochemistry, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Post Office Box 62000-00200, Nairobi, Kenya. [12]Department of Biochemistry, Virginia Polytechnic Institute and State University, 309 Fralin Hall, Blacksburg, VA 24061, USA. [13]Department of Biological Chemistry and Crop Protection, Rothamsted Research, West Common, Harpenden, Herts, AL5 2JQ, UK. [14]Department of Biological Sciences, McMicken College of Arts and Sciences, University of Cincinnati, Cincinnati, OH 45221, USA. [15]Department of Biological Sciences, University of Cape Town, Private Bag, Rondebosch, ZA-7700, South Africa. [16]Department of Biological Sciences, University of Wisconsin–Parkside, 900 Wood Road, Kenosha, WI 53144, USA. [17]Department of Biological Sciences, Wayne State University, 5047 Gullen Mall, Detroit, MI 48202, USA. [18]Department of Biology and Biotechnology, University of Pavia, Via Ferrata 9, Pavia, 27100, Italy. [19]Department of Biology, Baylor University, Waco, TX 76798, USA. [20]Department of Biology, KU Leuven, Naamsestraat 59, Leuven, B-3000, Belgium. [21]Department of Biology, New Mexico State University, Foster Hall 263, Las Cruces, NM 88003, USA. [22]Department of Biology, University of York, Wentworth Way, York, Y010 5DD, UK. [23]Department of Biology, West Virginia University, 53 Campus Drive, 5106 LSB, Morgantown, WV, USA. [24]Department of Biomedical Sciences, Institute of Tropical Medicine Antwerp, Nationalestraat 155, Antwerp, B-2000, Belgium. [25]Department of Cellular Biology, University of Georgia, 302 Biological Sciences Building, Athens, GA 30602, USA. [26]Department of Clinical Research, KEMRI-Wellcome Trust Programme, CGMRC, Post Office Box 230-80108, Kilifi, Kenya. [27]Department of Computer and Information Sciences, College of Science and Technology, Covenant University, P.M.B. 1023, Ota, Ogun State, Nigeria. [28]Department of Computer Science and Engineering, Department of Biochemistry and Biophysics, Texas A&M University, HRBB 328B TAMU, College Station, TX 77843, USA. [29]Department of Entomology, North Carolina State University, Campus Box 7613, Raleigh, NC 27695–7613, USA. [30]Department of Entomology, Texas A&M University, 2475 TAMU, College Station, TX 77843, USA. [31]Department of Entomology, The Ohio State University, 400 Aronoff Laboratory, 318 West 12th Avenue, Columbus, OH 43210, USA. [32]Department of Entomology, University of Arizona, 1140 East South Campus Drive, Forbes 410, Tucson, AZ 85721, USA. [33]Department of Entomology, University of Illinois at Urbana-Champaign, 505 South Goodwin Avenue, Urbana, IL 61801, USA. [34]Insect Pest Control Laboratory, Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture, Vienna, 1220, Austria. [35]Department of Environmental and Natural Resources Management, University of Patras, 2 Seferi Street, Agrinio, 30100, Greece. [36]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, 60 College Street, New Haven, CT 06520, USA. [37]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, 300 Aronoff Laboratory, 318 West 12th Avenue, Columbus, OH 43210, USA. [38]Department of Natural Sciences, St. Catharine College, 2735 Bardstown Road., St. Catharine, KY 40061, USA. [39]Department of Nutritional Sciences, University of Arizona, Career and Academic Services, College of Agriculture and Life Sciences, Forbes Building, Room 201, Post Office Box 210036, Tucson, AZ 85721–0036, USA. [40]Department of Nutritional Sciences, University of Arizona, Shantz 405, 1177 East 4th Street, Tucson, AZ 85721–0038, USA. [41]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK. [42]Department of Parasitology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK. [43]Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK. [44]Entomology Section, Onderstepoort Veterinary Institute, Private Bag X5, Onderstepoort, 110, South Africa. [45]Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA. [46]Department of Veterinary Tropical Diseases, University of Pretoria, Private Bag X04, Onderstepoort, 110, South Africa. [47]European Molecular Biology Laboratories, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, Cambridgeshire, CB10 1SA, UK. [48]Group of Bioinformatics and Modeling, Laboratory of Medical Parasitology, Biotechnology, and Biomolecules, Institut Pasteur de Tunis, 13, Place Pasteur, BP74, Belvédère, Tunis, 1002, Tunisia. [49]Institut de Recherche pour le Développement (IRD), UMR 177 IRD-CIRAD INTERTRYP, CIRDES Bobo-Dioulasso, Burkina Faso. [50]Institut de Recherche pour le Développement (IRD), UMR 177 IRD-CIRAD INTERTRYP, LRCT Campus International de Baillarguet, Montpellier, France. [51]Institute of Biological, Environmental, and Rural Sciences, University of Aberystwyth, Old College, King Street, Aberystwyth, Ceredigion, SY23 3FG, UK. [52]Institute of Biotechnology Research, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Post Office Box 62000-00200, Nairobi, Kenya. [53]Institute of Integrative Biology, The University of Liverpool, Crown Street, Liverpool, L69 7ZB, UK. [54]Institute of Zoology, Slovak Academy of Sciences, Dúbravská cesta 9, Bratislava, 845 06, Slovakia. [55]Integrated Database Team, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Aomi 2-4-7, Koto-ku, Tokyo, 135-0064, Japan. [56]Kenya Agricultural Research Institute Trypanosomiasis Research Centre, Post Office Box 362, Kikuyu, 902, Kenya. [57]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, 12735 Twinbrook Parkway, Room 2E-32D, Rockville, MD 20852, USA. [58]Technology Innovation Agency, National Genomics Platform, Post Office Box 30603, Mayville, Durban, 4058, South Africa. [59]Department of Biochemistry and Sports Science, Makerere University, Post Office Box 7062, Kampala, Uganda. [60]Bateman Group, Wellcome Trust Sanger Institute, EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Cambridgeshire, CB10 1SA, UK. [61]Molecular Biology and Bioinformatics Unit, International Center of Insect Physiology and Ecology, Duduville Campus, Kasarani, Post Office Box 30772-00100, Nairobi, Kenya. [62]National Livestock Resources Research Institute (NaLIRRI), Post Office Box 96, Tororo, Uganda. [63]National Research Center for Protozoan Diseases, Obihiro University of Agriculture and Veterinary Medicine, Inada-cho, Obihiro, Hokkaido, 080-8555, Japan. [64]Parasite Genomics Group, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Cambridgeshire, CB10 1SA, UK. [65]Riddiford Laboratory, Janelia Farm Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, VA 20147, USA. [66]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan. [67]South African National Bioinformatics Institute, South African MRC Bioinformatics Unit, University of the Western Cape, 5th Floor Life Sciences Building, Modderdam Road, Bellville 7530, South Africa. [68]Special Programme for Research and Training in Tropical Diseases (TDR), WHO, Avenue Appia 20, 1211 Geneva 27, Switzerland. [69]The Genome Institute, Washington University School of Medicine, St. Louis, MO 63110, USA. [70]Tsetse and Trypanosomiasis Research Institute (TTRI), Majani Mapana, Off Korogwe Road, Post Office Box 1026, Tanga, Tanzania. [71]Vector Health International, Post Office Box 15500, Arusha, Tanzania. [72]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Cambridgeshire, CB10 1SA, UK. [73]WHO Regional Office for Africa, WHO, Cité du Djoué, Post Office Box 06, Brazzaville, Congo. [74]School of Basic Sciences, Indian Institute of Technology, Mandi 175001, Himachal Pradesh, India. [75]Department of Parasite, Vector, and Human Biology, KEMRI-Wellcome Trust Programme, CGMRC, Post Office Box 230-80108, Kilifi, Kenya. [76]Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan. [77]Department of Biochemistry and Biotechnology, Kenyatta University, Post Office Box 43844-00100, Nairobi, Kenya. [78]Department of Biological Sciences, Indian Institute of Science Education and Research, Indore Bypass Road, Bhauri District, Bhopal, Madhya Pradesh, 462066, India.[79]Faculty of Science, King Abdulaziz University, Jeddah, 21589, SA. [80]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave. Boston, MA 02461. [81]Bioinformatics Research Unit, Agrogenomics Research Center, National Institute of Agrobiological Sciences, 2-1-2, Kannondai, Tsukuba, Ibaraki 305-8602, Japan.

Supplementary Materials

www.sciencemag.org/content/344/6182/380/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S9
Tables S1 to S43
References (*39*–*101*)

12 December 2013; accepted 19 March 2014
10.1126/science.1249656

Discovery of Brainwide Neural-Behavioral Maps via Multiscale Unsupervised Structure Learning

Joshua T. Vogelstein,[1,2]* Youngser Park,[1]* Tomoko Ohyama,[3]* Rex A. Kerr,[3] James W. Truman,[3] Carey E. Priebe,[1]†‡ Marta Zlatic[3]†‡

A single nervous system can generate many distinct motor patterns. Identifying which neurons and circuits control which behaviors has been a laborious piecemeal process, usually for one observer-defined behavior at a time. We present a fundamentally different approach to neuron-behavior mapping. We optogenetically activated 1054 identified neuron lines in *Drosophila* larvae and tracked the behavioral responses from 37,780 animals. Application of multiscale unsupervised structure learning methods to the behavioral data enabled us to identify 29 discrete, statistically distinguishable, observer-unbiased behavioral phenotypes. Mapping the neural lines to the behavior(s) they evoke provides a behavioral reference atlas for neuron subsets covering a large fraction of larval neurons. This atlas is a starting point for connectivity- and activity-mapping studies to further investigate the mechanisms by which neurons mediate diverse behaviors.

Nervous systems can generate a wide range of motor outputs, depending on their incoming sensory inputs and internal state. A comprehensive understanding of how behavioral diversity and selection is achieved requires the identification of neural circuits that mediate many distinct motor patterns in a given nervous system. Mapping a functional circuit for one be-

havior in a given organism is difficult enough, but doing so for many behaviors has been almost impossible. The first step in mapping a circuit that mediates a behavior is to identify neurons whose activity is causally related to the behavior. Such a list of neurons provides a starting point for identifying the connectivity patterns between the relevant neurons. Thus, to map circuits underlying many behaviors, one would need a comprehensive neuron-behavior atlas of the nervous system that would list all neurons causally related with each behavior.

Generating such neuron-behavior maps has been difficult for several reasons. First, the experimental tools to selectively manipulate small sets of neurons while simultaneously observing natural behavior were lacking. Fortunately, recent advances in genetic toolkits allow reasonably selective manipulation of neuron types in genetic model organisms, such as *Drosophila* (*1*–*3*). Advances in behavior tracking methods allow high-resolution monitoring of the effect of such manipulations (*4*, *5*). Neural manipulation screens can therefore be coupled with high-resolution monitoring of motor outputs to causally link complex behaviors to correspondingly complex neural circuits.

Second, establishing the causal links between neural manipulations and the resulting time-varying behavioral responses is a daunting computational statistics challenge. Existing supervised machine-learning methods can detect only predetermined behaviors (*6*); moreover, they are limited by the speed with which humans can annotate training data sets. An alternative approach uses unsupervised clustering of the multidimensional time series. However, the high-content and high-throughput nature of the time-varying behavior data presents both computational and statistical challenges.

We developed a methodology for data-driven neuron-behavior mapping and applied it to larval *Drosophila*. The nervous system of larval *Drosophila* consists of a well-developed brain and nerve cord containing only about 10,000 neurons, rendering it sufficiently simple to obtain a relatively comprehensive characterization of it. Moreover, there exist more than 1000 genetic GAL4 lines in *Drosophila* larvae with recently characterized sparse neuronal expression patterns that together cover most of the 10,000 neurons in the larval nervous system (http://flweb.janelia.org/cgi-bin/flew.cgi) (*3*).

Optogenetic Neural Activation Screen

We designed an optogenetic neural activation screen (see supplementary materials) to obtain a neuron line–behavior atlas of the larval nervous system that would contain causal links between neuron lines and the motor patterns they control. We used 1049 distinct GAL4 lines to selectively target channelrhodopsin-2 (ChR2) (*7*) to sparse distinct subsets of neurons, with each line activating 2 to ~15 neurons. Because these lines essentially span the entire set of larval neurons, some lines activate sensory and motor neurons as well as many neurons involved in decisions and action selection. We included four positive control lines in the screen that drive expression in nociceptive, mechanosensory, and proprioceptive neurons, previously determined to reliably mediate distinct behaviors (*8*–*10*), as well as one negative control line in which no neurons were optogenetically activated (*2*, *10*), for a total of 1054 lines and 37,780 animals tested. In each experiment, we exposed dishes of larvae to 470-nm light stimuli (one exposure of 30 s followed by four exposures of 5 s, with a 30-s interval after the long exposure and a 10-s interval between the short exposures) to optogenetically activate Ch2-expressing neurons; we captured video before, during, and after stimulation (Fig. 1A). The Multi-Worm Tracker (MWT) software (*4*) tracked time-varying, two-dimensional closed contours of larvae and sketched eight time-varying features that collectively characterize larval shape and motion (Fig. 1B). Streaming and sketching reduced the data complexity by a factor of more than 200,000, enabling a compressive yet expressive representation of the data. These reduced data served as the input into the multiscale unsupervised structure learning methodology to reveal data-driven behavior types (Fig. 1C). Each behavior type was then linked to the subset of lines that mediate them (Fig. 1D).

Discovery of Behavior Types via Multiscale Unsupervised Structure Learning

As a first step, we sought to discover a large, inclusive, and nonpredetermined set of statistically distinguishable behavioral responses performed by the 37,780 animals during the first (30-s) optogenetic activation period. Recently developed methods for multiscale unsupervised structure learning (*11*–*13*) can be thought of as generalizations of manifold learning techniques, in that they can learn structures more general than manifolds, such as unions of manifolds. We adopted iterative denoising tree (IDT) methodology (*11*, *14*), which offers demonstrated utility across several domains (*15*, *16*).

The input to IDT is the collection of all 37,780 larval sketches, irrespective of which line generated each sketch (Fig. 2A). IDT consists of five key

[1]Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21218, USA. [2]Department of Statistical Science, Duke University, Durham, NC 27708, USA. [3]Janelia Farm Research Campus, Ashburn, VA 20147, USA.

*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. E-mail: cep@jhu.edu (C.E.P.); zlaticm@janelia.hhmi.org (M.Z.)

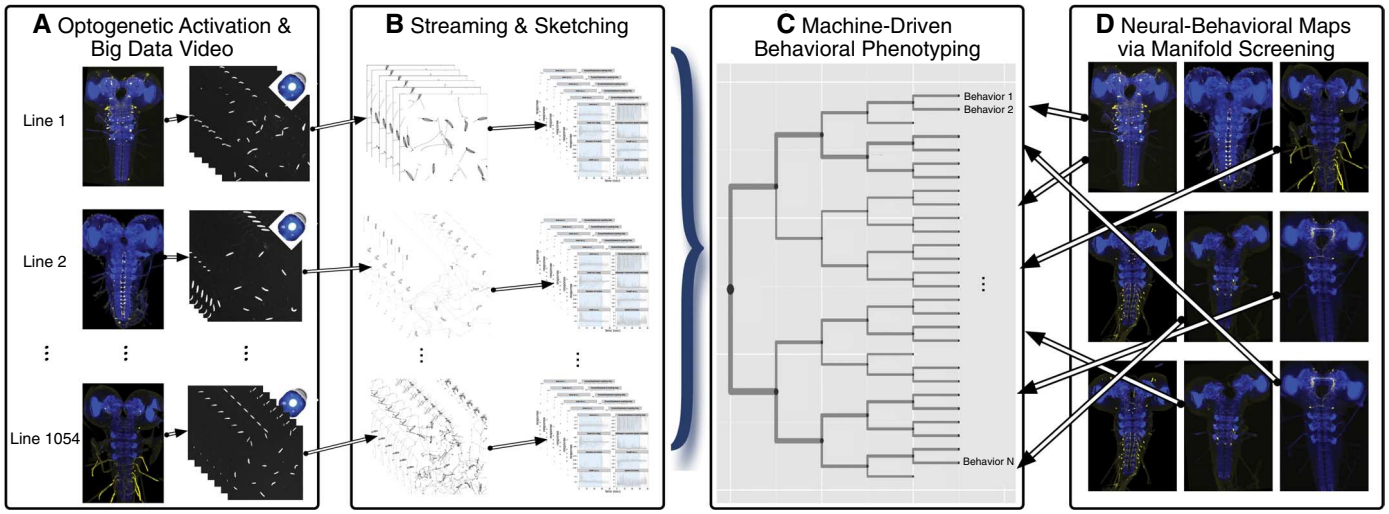

Fig. 1. Experimental design and methodology for obtaining neuron line–behavior maps. (**A**) Optogenetic activation screen of 1054 lines while digitally recording high-dimensional larval responses. (**B**) Streaming extracts the contours of each larva from each video frame; sketching extracts eight time-varying features from the contours that characterize the shape and motion of each animal. (**C**) Machine-driven behavioral phenotyping learns phenotype categories (called behaviotypes) from the sketches via multiscale unsupervised structure learning. (**D**) Manifold testing discovers which neuron lines evoke sets of behaviors that are different from negative controls, which facilitates associating each such line with some number of behaviotypes.

steps collectively resulting in a hierarchical clustering tree. In step one, IDT computes a dissimilarity between all pairs of sketches (Fig. 2B). The dissimilarity choice can have drastic computational and inferential impact; thus, choosing one that captures signal variability, and can be efficiently computed, is key to the success of IDT. This is important because the 37,780 observations yield more than 1.4 billion dissimilarities. We used a smoothed distance between the sketches of the data (*17*) as a dissimilarity function (see supplementary materials).

In step two, IDT transforms the interpoint dissimilarity matrix into a set of n relatively high-dimensional Euclidean vectors, via an approach such as multidimensional scaling (*18*). Once each data point is in Euclidean space, an extensive toolkit from classical statistical machine learning is applicable (*19*).

Step three consists of iterating, once per tree depth, the following substeps: (i) select a subset of the dimensions from each cluster, (ii) cluster all the nodes of the partition tree obtained thus far, and (iii) check for convergence at each resulting cluster. For each cluster at each scale, the number of dimensions to select and the number of subclusters to generate are decided in a data-driven adaptive fashion. The final result is a hierarchical tree characterizing families and subfamilies of behavioral responses (Fig. 2C).

To visualize the phenotype categories learned by IDT (or nodes/clusters of the tree), called behaviotypes, Fig. 2D shows, for each cluster, the time-varying means and standard errors for each of the eight sketched time-varying features. Each behaviotype is well separated from at least some other cluster for at least some of the time for at least some features. This provides an intuitive validation of the uniqueness of the behaviotypes, thereby demonstrating the efficacy of IDT.

IDT Identified Both Previously Described and Novel Behavioral Sequences

Inspection of the time-varying means of the responses of each behaviotype cluster (Fig. 2D) and of the videos of animals at the center of each cluster (movies S1 to S58) revealed that IDT clustered many of the larval behavioral responses into categories similar to those a human expert would have identified. We could label the nodes of the learned tree post hoc (Fig. 2, E and F). For example, the first division revealed by IDT differentiates between slow and fast families. The fast family is subdivided into turn-avoid and escape-crawl subfamilies. IDT further distinguished between the right turn–left turn–avoid and the symmetric left turn–right turn–avoid sequence (fig. S1) and between two types of escape crawl that are preceded by more or less hunching and wiggling (fig. S2). The slow family is subdivided into still or backup and turner; the still or backup is further subdivided into still (no movement) and backup (lots of backward crawling) (fig. S3), and the turner is subdivided into turn-turn-turn (continuous turning) and turn-slow-crawl (fig. S4).

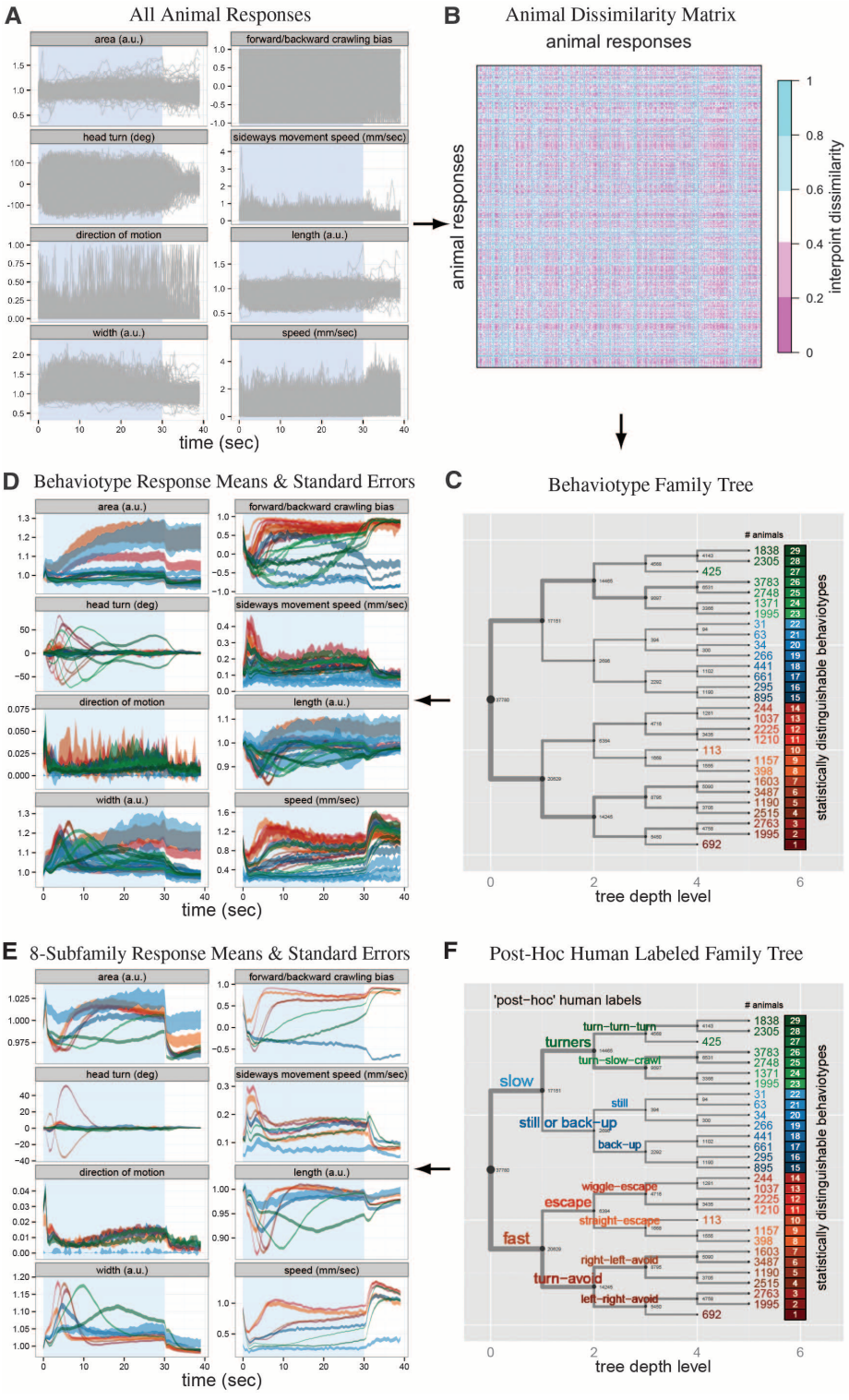

Fig. 2. IDT detected 29 distinguishable behaviotypes. (**A**) The input to IDT is the collection of all 37,780 larval sketches. Structure is not obvious. Blue shading represents the period of photostimulation in all figures. (**B**) Dissimilarity matrix between all pairs of animals. Only ~1.4 million of the ~1.4 billion pairwise dissimilarities are shown. (**C**) $\hat{K}$ = 29 distinguishable behavior clusters (behaviotypes) were identified using IDT to learn a data-adaptive clustering of the high-content responses. (**D**) Mean and standard error of the responses of each of the 29 behaviotypes identified in (C), using the same color code as in (C). (**E**) Mean and standard error of the responses of the eight behaviotype subfamilies. (**F**) The same tree as in (C) but with post hoc human labels assigned to the automatically detected behaviotype families and subfamilies.

After level three of the behaviotype tree, subjective visual inspection of the videos failed to detect differences between related behaviotypes. Nonetheless, they do have distinct properties for some of the features for some of the time (Fig. 2D). Behaviotypes 17 and 18 represent animals with tracking errors in which the MWT streaming and sketching software inverted the front and back of the animals (fig. S6 and movies S33 to S36). IDT assigned animals with this type of tracking error to the two separate behavioral clusters, conveniently isolating such errors.

Many of the optogenetically evoked, automatically detected behaviotype subfamilies are similar to previously described larval responses to various natural stimuli (*10*, *20*, *21*). IDT also detected behavioral categories not previously documented. This is unsurprising because only

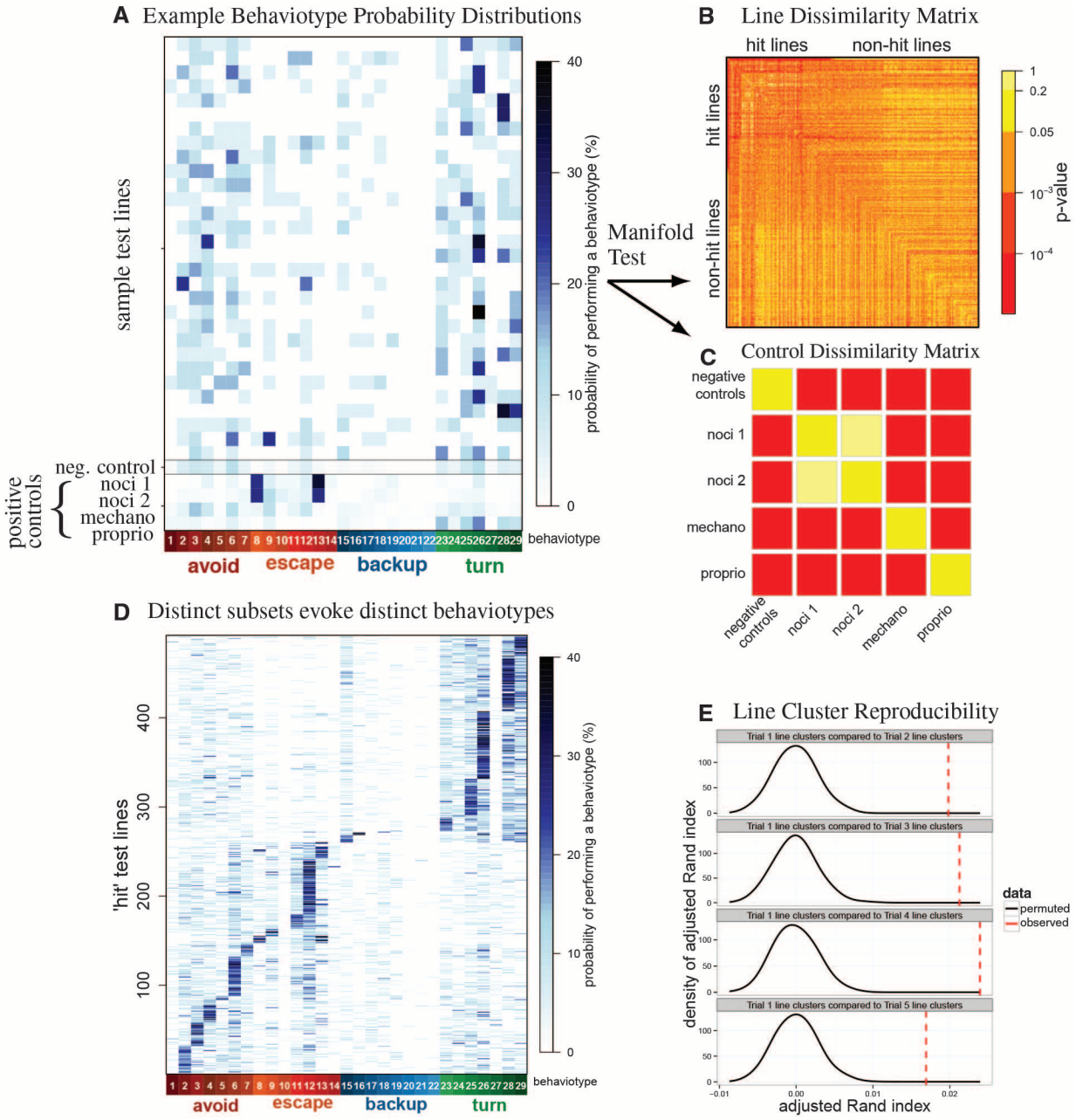

Fig. 3. Learning neuron line–behaviotype maps via manifold testing. (**A**) Behaviotype probability vectors for a number of lines. Each row shows the percentage of animals performing a behaviotype for optogenetic activation of a particular line: 30 randomly sampled test lines (top), negative controls (pBDPU-ChR2) with no optogenetically activated neurons (middle), and positive controls—nociceptive (ppk-ChR2, noci 1; R38A10-ChR2, noci 2); mechanosensory (iav-ChR2); and proprioceptive (R11F05-ChR2) neuron lines with known effects on behavior (bottom). (**B**) Line dissimilarity matrix showing pairwise *P* values for all 1054 lines computed via the manifold test. The first entry is the negative control. Remaining entries are sorted according to *P* value for the comparison with the negative control, from lowest to highest; 455 lines (4 positive controls + 451 test lines) were significantly different from the negative control (hit lines), and 598 lines were not (nonhit lines). (**C**) The *P* values between all pairs of known somatosensory neuron lines and negative controls are "correct": Those that should be significantly different according to previous studies are, and those that should not be are not. This panel is a subset (top left 5 × 5 submatrix entries) of (B). (**D**) Behaviotype probability distributions for all 451 significantly distinct test neuron lines, sorted according to their maximum probability behaviotype. (**E**) The clusters of lines that bias the probability toward the same behaviotypes are reproducible, demonstrated by the fact that the mode of the empirical null distribution of the adjusted Rand index (ARI) for reliable clustering (black curve) is lower in all four repeated trials than the observed ARI (all *P* values < 0.01).

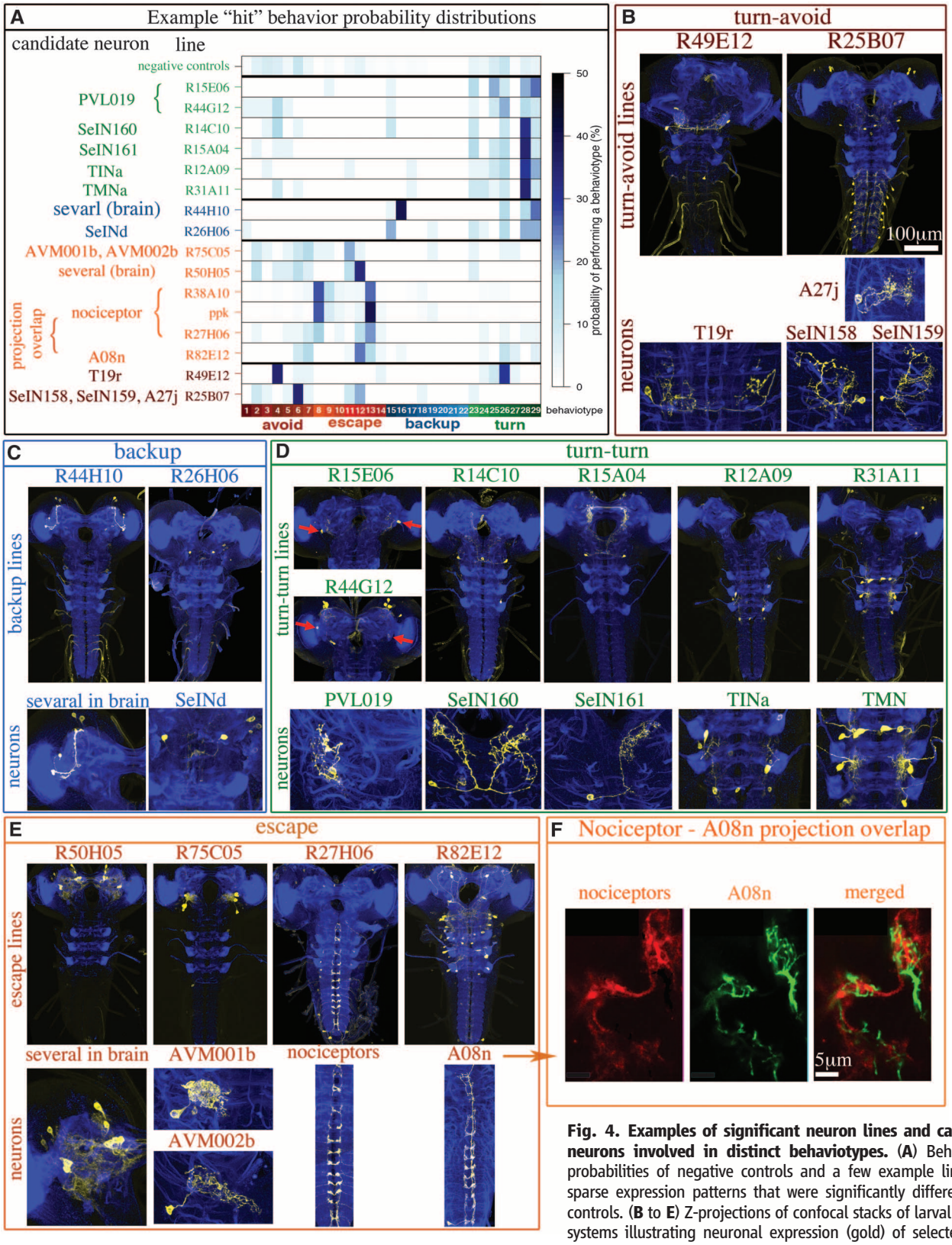

Fig. 4. Examples of significant neuron lines and candidate neurons involved in distinct behaviotypes. (**A**) Behaviotype probabilities of negative controls and a few example lines with sparse expression patterns that were significantly different from controls. (**B** to **E**) Z-projections of confocal stacks of larval nervous systems illustrating neuronal expression (gold) of selected lines. Neuropil is stained for reference with antibody to N-cadherin (blue). High-resolution images of candidate neurons from these lines are shown beneath the lower-magnification images of the entire nervous system. These lines biased the probability toward (B) turn-avoid, (C) backup, (D) turn-turn, and (E) escape behaviotype subfamilies. In some cases, two or more lines that drive in the same neuron biased the probability toward the same behaviotype. R15E06 and R44G12 both drive in PVL019 (red arrows) and biased the probability toward turn-turn behaviotypes 25, 26, 28, and 29. Ppk, R38A10, and R27H06 drive in the same nociceptor neurons and biased the probability toward escape behaviotype 13. (**F**) R82E12 projections (green) overlap tightly with nociceptive axon terminals (red) in a single 0.5-µm confocal section, likely forming synaptic contacts. These two lines both biased the probability toward the same escape behaviotype 13 and likely belong to the same anatomical and functional neural circuit.

a relatively small portion of the larval stimulus space has been explored in previous studies.

Biased Probabilistic Relationship Between Neuron Activation and Behaviotypes

We used the prior results that nociceptive, mechanosensory, and proprioceptive sensory neurons reliably mediate distinct behaviors to assess the validity of this behavior discovery approach. The optogenetic screen included lines that targeted ChR2 expression to the nociceptive [ppk and R38A10 (*10*, *22*)], mechanosensory (iav) (*9*), and proprioceptive (R11F05) neurons. Activating distinct sensory neurons mostly biased the behaviotype probability toward distinct subfamilies: (i) Nociceptive stimulation tended to evoke escape behaviotypes; (ii) mechanosensory stimulation tended to evoke turn-avoid behaviotypes; and (iii) proprioceptive tended to evoke slow behaviotypes (Fig. 3A). In contrast, activating the same nociceptive neurons with two distinct GAL4 lines evoked similar behaviotype probability distributions. The response profiles of the negative controls (pBDPU-ChR2) (*2*, *10*) that did not express ChR2 likely represented larval reactions to blue light alone and appeared different from those of animals with optogenetically activated nociceptive, mechanosensory, or proprioceptive neurons (Fig. 3A). Finally, analysis of the 1049 previously unexplored neuronal test lines (Fig. 3A; top rows show 30 examples) demonstrated that distinct lines bias the probability toward distinct behaviotypes. Optogenetic stimulation of different animals of the same line (that is, activating the "same" neurons in different animals) did not always evoke the same behavior; rather, it biased the probability toward a few possible behaviotypes.

We also applied IDT jointly to the responses from the four additional identical repeats of the 5-s optogenetic activation stimulus of the same organisms and analyzed the behaviotype distributions of the same individuals on different trials (fig. S5A). We found that even repeated activation of the same neurons in the same individual did not always evoke the same behaviotype; rather, it biased the probability toward a few possible behaviotypes (fig. S5A). However, the responses of individual animals were significantly more similar to each other than to the responses of distinct individuals (fig. S5B).

Screening Neuron Lines via Manifold Testing

Each neuron line is characterized by the empirical probability of larvae performing each behaviotype and is encoded by a $\hat{K} = 29$-dimensional vector with nonnegative entries that sum to unity. This encoding enables direct testing of each pair of lines by choosing an appropriate test statistic and applying a standard test. However, the choice of test statistic for these multivariate probability vectors is not obvious; moreover, existing tests do not sufficiently address the multiple dependent hypothesis–testing problem. We therefore devised the following manifold test (see supplementary materials).

The main idea of the manifold test is to jointly and nonlinearly embed each experiment into a lower-dimensional representation. The test first computes the Hellinger distance between all 1054^2 pairs of experiments and then uses multidimensional scaling (MDS) (*18*) to obtain a low-dimensional Euclidean embedding. In this space, one can easily compute the difference between any pair of trials via an "out-of-sample" extension to this embedding methodology (*23*). Moreover, via bootstrap, the test computes the significance of those differences (*24*). Then it adjusts the *P* values to account for multiple correlated hypotheses (*25*) and other batch effects (*26*). Using this manifold test, we identified a large fraction of lines (4 positive controls and 451 test lines) as being significantly different (hit lines) from the negative control line (Fig. 3B; see supplementary data set 1 for the list of all such lines).

The positive control lines with different neurons activated (nociceptive, mechanosensory, and proprioceptive) were all significantly distinct from the negative controls and from each other (Fig. 3C). The two positive control lines with the same nociceptive neurons activated (ppk and R38A10) were not significantly different from one another (Fig. 3C). The known somatosensory neuron controls lend additional credence to this manifold test.

A Neuron Line–Behaviotype Atlas

Given the list of lines significantly distinct from the negative controls, we desired to identify which lines bias the probability toward which behaviotypes, thereby generating a neuron line–behaviotype atlas. Many lines biased the probability of behavior primarily to one behaviotype, with the remaining probability distributed to a few related behaviotypes (Fig. 3D).We confirmed that the line clusterings are reliable by applying IDT independently to the four additional repeats of the 5-s optogenetic activation stimulus of the same organisms, and then comparing how similar the line clusters from each of these additional trials were to the first trial. The identified line clusters were indeed reproducible ($P < 0.001$ via an adjusted Rand index permutation test, Fig. 3E).

For most behaviotypes, activation of at least one line biased the probability toward that behaviotype significantly more than the negative controls (see supplementary data set 1 for significant line–behaviotype probability distribution numbers and *P* values). Collectively, these significance results constitute a reference atlas that associates each neural line with a set of behaviors it mediates (and vice versa, associating each behaviotype with a set of neural lines that mediates it). Images of neuronal expression patterns of all of these lines are available from http://flweb.janelia.org/cgi-bin/flew.cgi.

Figure 4 shows examples of behaviotype probability vectors (Fig. 4A) and neuronal expression patterns of a few significant lines that bias the probability toward turn-avoid (Fig. 4B), backup (Fig. 4C), turn-turn (Fig. 4D), or escape subfamilies (Fig. 4E). We analyzed with higher resolution the projections of individual neurons from some of these lines using a single-cell flip-out method (*27*) and gave names to the identified neurons. A number of the significant lines only drove expression in single pairs of larval neurons, yet they still biased the probability toward specific behaviotypes. For example, activation of the pair of PVL019 neurons in the basal posterior region of the larval brain (with the line R15E06) or of the pair of SeIN161 neurons in the subesophageal ganglion (with the line R15A04) evoked turn-turn-turn behaviotypes in a large fraction of animals. This indicates that single pairs of cells can exert drastic control over behavior, consistent with findings in other systems (*28*, *29*).

In some cases, we found that lines that contribute to the same behaviotypes contain neurons with anatomically overlapping projections that may constitute presynaptic and postsynaptic partners. For example, the motor neurons in R31A11 and the interneurons in R12A09 both strongly evoked turn-turn-turn behaviotype 28, and their projections occupy the same motor region of the nerve cord and likely overlap. Similarly, R27H06 and R82E12 (Fig. 4E) biased the probability toward escape behaviotype 13 and drive expression in the nociceptive neurons, and in a class of ascending projection neurons (A08n), respectively. Double labeling of the nociceptive and the A08n neurons confirmed that their arbors do indeed tightly overlap, likely forming synaptic contacts (Fig. 4F).

Discussion

An important step toward understanding the structure and function of neural circuits is to identify comprehensive lists of neurons whose activation is causally related to a comprehensive set of motor patterns. This requires statistical approaches for identifying structure in high-dimensional behavior data.

Methods for systematic, automated, and unsupervised clustering are rarely applied to high-dimensional behavioral data. A recent study used unsupervised learning to detect behavioral motifs defined as sequences of four "eigenworm" positions in mutant and wild-type freely moving *Caenorhabditis elegans* in the absence of any stimulation (*30*). Here, we developed methods for clustering entire behavioral response sequences during a stimulation period and made use of discovered behavioral clusters to detect significantly distinct experiments from a large-scale screen, providing statistical validation that the learned clusters are meaningful. We applied these methods to discover neurons causally related to behaviors, providing a proof of principle that it is possible to classify neurons in terms of their activation of motor outputs in a fully automated way.

The collection of neuron line–behaviotype relationships collectively constitutes an atlas. Starting with a list of lines whose activation is sufficient to evoke each behaviotype, the relevant

individual neurons from each line can readily be identified. Whereas some lines drive expression in a single pair of neurons, most drive in the range of two to five candidate neuron types. In cases where lines drive in more than one neuron, intersectional strategies can be used to target individual neurons and test the effect of their activation on behavior (*2*).

This reference atlas provides a valuable starting point for understanding how distinct behaviors are selected and controlled. Large-scale connectomics (*31*–*33*) and functional brain imaging methods (*34*, *35*) will soon provide similarly comprehensive views of the structure of neural circuits and of the activity patterns within those circuits. However, a connectome by itself does not carry information about which neurons mediate which behaviors. Similarly, a brain-activity map alone shows the flow of information through the network, but does not reveal causal relationships between neurons and behavior. Together, the neuron-behavior map, the neuron-activity map, and the connectome complement one another, laying the groundwork for a brainwide understanding of the principles by which brains generate behavior.

The statistical methods presented here are generally applicable to discovery of scientifically meaningful structure from big data—a pressing problem in the information age.

References and Notes

1. B. D. Pfeiffer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9715–9720 (2008).
2. B. D. Pfeiffer *et al.*, *Genetics* **186**, 735–755 (2010).
3. A. Jenett *et al.*, *Cell Rep.* **2**, 991–1001 (2012).
4. N. A. Swierczek, A. C. Giles, C. H. Rankin, R. A. Kerr, *Nat. Methods* **8**, 592–598 (2011).
5. K. Branson, A. A. Robie, J. Bender, P. Perona, M. H. Dickinson, *Nat. Methods* **6**, 451–457 (2009).
6. M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, K. Branson, *Nat. Methods* **10**, 64–67 (2013).
7. C. Schroll *et al.*, *Curr. Biol.* **16**, 1741–1747 (2006).
8. R. Y. Hwang *et al.*, *Curr. Biol.* **17**, 2105–2116 (2007).
9. C. W.-H. Wu *et al.*, *Neuron* **70**, 229–243 (2011).
10. T. Ohyama *et al.*, *PLOS ONE* **8**, e71706 (2013).
11. C. E. Priebe, D. J. Marchette, D. M. Healy, *Mod. Signal Process.* **46**, 223 (2003).
12. W. K. Allard, G. Chen, M. Maggioni, *Appl. Comput. Harmon. Anal.* **32**, 435–462 (2012).
13. P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer, D. Morozov, in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)* (IEEE, Providence, RI, 2007), pp. 536–546.
14. K. E. Giles, M. W. Trosset, D. J. Marchette, C. E. Priebe, *Comput. Stat.* **23**, 497–517 (2008).
15. C. E. Priebe, D. J. Marchette, D. M. Healy, *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 699–708 (2004).
16. D. Karakos, S. Khudanpur, J. Eisner, C. E. Priebe, in *Proceedings of the 2005 IEEE International Conference on Acoustics Speech and Signal Processing ICASSP* (IEEE, Philadelphia, 2005), vol. 5, pp. 1081–1084.
17. C. E. Priebe, *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 404–413 (2001).
18. I. Borg, P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer, New York, 2010).
19. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2001).
20. E. A. Kane *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, E3868–E3877 (2013).
21. M. Kernan, D. Cowan, C. Zuker, *Neuron* **12**, 1195–1206 (1994).
22. J. A. Ainsley *et al.*, *Curr. Biol.* **13**, 1557–1563 (2003).
23. M. Tang, Y. Park, C. E. Priebe, http://arxiv.org/abs/1305.4893 (2013).
24. P. I. Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses* (Springer, New York, 2010).
25. D. Yekutieli, Y. Benjamini, *Ann. Stat.* **29**, 1165–1188 (2001).
26. R. A. Fisher, *Statistical Methods for Research Workers* (Oliver and Boyd, Edinburgh, 1925).
27. G. Struhl, K. Basler, *Cell* **72**, 527–540 (1993).
28. I. Kupfermann, K. R. Weiss, *Behav. Brain Sci.* **1**, 3–10 (1978).
29. B. Hedwig, *J. Neurophysiol.* **83**, 712–722 (2000).
30. A. E. X. Brown, E. I. Yemini, L. J. Grundy, T. Jucikas, W. R. Schafer, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 791–796 (2013).
31. D. D. Bock *et al.*, *Nature* **471**, 177–182 (2011).
32. M. Helmstaedter, *Nat. Methods* **10**, 501–507 (2013).
33. S.-Y. Takemura *et al.*, *Nature* **500**, 175–181 (2013).
34. M. B. Ahrens, M. B. Orger, D. N. Robson, J. M. Li, P. J. Keller, *Nat. Methods* **10**, 413–420 (2013).
35. T. Schrödel, R. Prevedel, K. Aumayr, M. Zimmer, A. Vaziri, *Nat. Methods* **10**, 1013–1020 (2013).

Acknowledgments: We thank G. M. Rubin, B. D. Pfeiffer, A. Nern, and B. Condron for fly stocks; B. D. Mensh for exceptionally helpful comments on the manuscript; A. Cardona, J. H. Simpson, and K. Branson for helpful discussions; C. Sullivan and A. Mondal for help with editing; H. Li and Fly Light Project Team at Janelia HHMI for images of neuronal lines; Janelia Fly Core for setting up the fly crosses for the activation screen; and Janelia Scientific Computing for help with data processing and storage, especially E. Trautman, R. Svirskas, and D. Olbris. Supported by the Larval Olympiad Project and Janelia HHMI, the XDATA program of the Defense Advanced Research Projects Agency administered through Air Force Research Laboratory contract FA8750-12-2-0303, and a National Security Science and Engineering Faculty Fellowship. All raw data, data derivatives, and code are freely available from http://openconnecto.me/behaviotypes.

Supplementary Materials

www.sciencemag.org/content/344/6182/386/suppl/DC1
Materials and Methods
Figs. S1 to S6
Movies S1 to S58
Supplementary Data Sets 1 and 2
References (*36*–*40*)

31 December 2013; accepted 17 March 2014
Published online 27 March 2014;
10.1126/science.1250298

REPORTS

A Dual-Catalysis Approach to Enantioselective [2 + 2] Photocycloadditions Using Visible Light

Juana Du,* Kazimer L. Skubi,* Danielle M. Schultz,* Tehshik P. Yoon†

In contrast to the wealth of catalytic systems that are available to control the stereochemistry of thermally promoted cycloadditions, few similarly effective methods exist for the stereocontrol of photochemical cycloadditions. A major unsolved challenge in the design of enantioselective catalytic photocycloaddition reactions has been the difficulty of controlling racemic background reactions that occur by direct photoexcitation of substrates while unbound to catalyst. Here, we describe a strategy for eliminating the racemic background reaction in asymmetric [2 + 2] photocycloadditions of α,β-unsaturated ketones to the corresponding cyclobutanes by using a dual-catalyst system consisting of a visible light–absorbing transition-metal photocatalyst and a stereocontrolling Lewis acid cocatalyst. The independence of these two catalysts enables broader scope, greater stereochemical flexibility, and better efficiency than previously reported methods for enantioselective photochemical cycloadditions.

Modern stereoselective synthesis enables the construction of a vast array of organic molecules with precise control over their three-dimensional structure (*1*, *2*), which is important in a variety of fields ranging from drug discovery to materials engineering. Photochemical reactions could have a substantial impact on these fields by affording direct access to certain structural motifs that are otherwise difficult to construct (*3*, *4*). For example, the most straightforward methods for the construction of cyclobutanes and other strained four-membered rings are photochemical [2 + 2] cycloaddition reactions. The stereochemical control of photocycloadditions, however, remains much more challenging than the stereocontrol of analogous non-photochemical reactions (*5*, *6*) despite the chemistry community's sustained interest in photochemical stereoinduction over the last century (*7*, *8*).

Although many strategies using covalent chiral auxiliaries (*9*, *10*) or noncovalent chiral controllers (*11*, *12*) have been used to dictate absolute stereochemistry in photochemical cycloaddition reactions, the development of methods that utilize substoichiometric stereodifferentiating chiral catalysts has proven a more formidable challenge.

Department of Chemistry, University of Wisconsin–Madison, 1101 University Avenue, Madison, WI 53706, USA.

*These authors contributed equally to this work.
†Corresponding author. E-mail: tyoon@chem.wisc.edu

Discovery of Brainwide Neural-Behavioral Maps via Multiscale Unsupervised Structure Learning

Joshua T. Vogelstein, Youngser Park, Tomoko Ohyama, Rex A. Kerr, James W. Truman, Carey E. Priebe and Marta Zlatic

Science **344** (6182), 386-392.
DOI: 10.1126/science.1250298originally published online March 27, 2014

Optogenetic Insights

Mapping functional neural circuits for many behaviors has been almost impossible, so **Vogelstein *et al.*** (p. 386, published online 27 March; see the Perspective by **O'Leary and Marder**) developed a broadly applicable optogenetic method for neuron-behavior mapping and used it to phenotype larval *Drosophila* and thus developed a reference atlas. As optogenetic experiments become routine in certain fields of neuroscience research, creating even more specialized tools is imperative (see the Perspective by **Hayashi**). By engineering channelrhodopsin, **Wietek *et al.*** (p. 409, published online 27 March) and **Berndt *et al.*** (p. 420) created two different light-gated anion channels to block action potential generation during synaptic stimulation or depolarizing current injections. These new tools not only improve understanding of channelrhodopsins but also provide a way to silence cells.

CellPress

Neuron
NeuroView

To the Cloud! A Grassroots Proposal to Accelerate Brain Science Discovery

Neuro Cloud Consortium*
*Correspondence: jovo@jhu.edu (Joshua T. Vogelstein)
http://dx.doi.org/10.1016/j.neuron.2016.10.033

The revolution in neuroscientific data acquisition is creating an analysis challenge. We propose leveraging cloud-computing technologies to enable large-scale neurodata storing, exploring, analyzing, and modeling. This utility will empower scientists globally to generate and test theories of brain function and dysfunction.

Introduction

Technological advances from all around the globe (Grillner et al., 2016) are allowing neuroscientists to collect more precise, complex, varied, and extensive data than ever before (Sejnowski et al., 2014). How can we maximally accelerate our collective ability to extract meaning from such data? To answer this question, the United States Congress commissioned the National Science Foundation (NSF) to "convene government representatives, neuroscience researchers, private entities, and non-profit institutions" (https://www.congress.gov/congressional-report/113th-congress/house-report/448). The NSF funded two events. The first was a workshop of over 75 individuals from 12 countries and 5 continents that was broadcast live over the internet. Each person was invited to bring a single big idea—one that could have maximal impact, while being both feasible, given existing resources, and universally inclusive. Four ideas emerged as grand challenges for global brain science (Vogelstein et al., 2016). A second event was organized to discuss these ideas with a larger (425 participants) and more diverse community, which will be the subject of another article. The goal of this NeuroView is to describe one of the four grand challenges and propose a strategy to overcome it, in order to gather feedback from the larger community. The authors are participants in the first conference who volunteered to hash out these ideas via emails, online documents, conference calls, and in-person visits.

The kernel of the idea is based on a view of the scientific process as an "upward spiral": a collective effort where each new experiment yields data, upon which analysis is performed, leading to new or refined models, which suggest novel experiments (see Figure 1). Historically, the process of data analysis has been kept relatively simple by the small scale of data acquired. But recent advances in experimental technology, such as serial electron microscopy (Denk and Horstmann, 2004), light sheet microscopy (Weber et al., 2014), and models of the whole human brain at the microscopic level (Amunts et al., 2013), have made data analysis significantly more challenging. While experimental neuroscience is enabling the collection of ever larger and more varied datasets, information technology is undergoing a revolution of its own. Commercial development of artificial intelligence and cloud computing innovations are changing the computational landscape (The Economist, 2016). Computing is moving toward "cloudification," a "software as a service" model, in which locally installed software programs are replaced by web apps. These forces create a massive opportunity to develop new computational technologies that complement advances in data collection in order to accelerate and democratize model building, hypothesis testing, and model refinement.

What Would Change If We Capitalize on This Opportunity?

Consider sending a letter, watching a movie at home, or obtaining reference information. Ten to twenty years ago, to send a letter, we purchased paper, stamps, and envelopes; to watch a movie at home, we rented or purchased a VHS or DVD; to obtain reference information, we bought an encyclopedia and obtained yearly revisions. Today, each of those options is still available and indeed preferred in certain circumstances. However, web options exist for each activity as well. In each case, we have privacy, bandwidth, and financial concerns. Nonetheless, for many of our daily practices we use these cyber solutions, sometimes putting our most private information in the cloud. The everyday practice of brain science is just beginning to benefit from similar technology development.

Other scientific disciplines have already navigated similar waters with remarkable success. For example, the Sloan Digital Sky Survey (SDSS) changed the daily practice of astronomers and cosmologists (Kent, 1994). They still have the option to wait 6 months for telescope time, analyze their data locally on machines they own and maintain, and publish a summary of the results (and many do). Yet there are more accounts in SDSS than there are professional cosmologists. Astronomers can now log in to SDSS, find previously published data, run database queries (a skill they typically did not have prior to SDSS), and publish the queries and results. Similarly, molecular geneticists historically sequenced their own data (using machines that they owned and maintained), analyzed it locally, and published the results. Now, they can outsource the sequencing to avoid owning and maintaining the machines, upload the sequences to a national or international database, quantitatively compare their sequences to previously published sequences, and then publish their findings. The success of these efforts is evident from the cultural shift of daily practices by many, if not most, participants in each field. Both fields resolved issues of data privacy, data ownership, governance, and financial concerns, providing a proof of principle that other scientific disciplines can do the same.

In neuroscience, many of our scientific practices remain based on pre-internet methods. A scientist designs an

experiment, collects data, stores it locally, keeps metadata in his head or in some custom spreadsheet, analyzes it using software that he buys and installs on local computers that he updates regularly, and publishes a summary of the results. We predict that another strategy will be superior for many situations: as the scientist collects data, it gets stored privately or publicly in the cloud, and she then selects analyses to occur automatically, having the flexibility to pull from a variety of previously published analyses, and finally publishes entire "digital experiments," containing (some of) the data and the entire analysis pipeline.

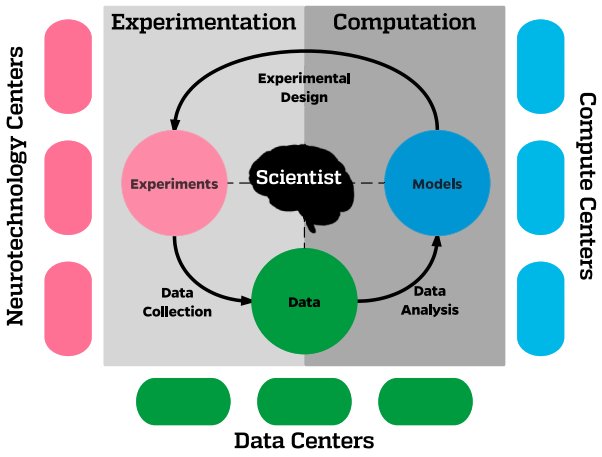

Figure 1. The Upward Spiral of Science

What Are the Primary Goals?

We see two key goals that, if achieved, would leverage advances in computing to accelerate brain sciences. The first goal is to make reproducibility and extensibility of science as easy as possible, even for small amounts of data or simple data. The current practices of private data storage and siloed analyses make reproducing an analytic result tedious at best and impossible at worst. The steps can include requesting the data, identifying the formats and organization, requesting the code, deciding which functions to run and how, getting all necessary dependencies installed, making sure to use the same software versions, and accessing the same computational hardware. Solutions now exist to mitigate each of these challenges, though they are relatively disparate and unconnected. Data can be uploaded to data repositories (e.g., https://figshare.com/), data standards have been proposed for several domains of brain science (e.g., http://bids.neuroimaging.io/ and http://www.nwb.org/), code can be stored in publicly accessible repositories (e.g., https://github.com/), interactive tutorials can be provided (e.g., using http://jupyter.org/), and all necessary software dependencies can be easily packaged together (e.g., using https://www.docker.com/) and run "in the cloud" (e.g., using http://mybinder.org/) on commercial service providers (e.g., on https://aws.amazon.com/ec2/ or https://cloud.google.com/). Nonetheless, given some new data, it is not obvious where to find reference algorithms or how to connect them to the data. Similarly, given a new model, it is not clear how to find reference data, figure out which standard it is using and then fit it, and determine if others have done the same to allow us to compare and assess the results. In either case, once the data are processed, it remains difficult to keep track of the resulting data derivatives and which version of which code resulted in which outputs. So although many of the pieces are in place, there is still no unified "glue" that makes everything work together seamlessly. Moreover, each of the above-mentioned tools can be used by some brain scientists, but most tools are designed for data scientists, so the learning curve can be incredibly steep. Ideally, there would be a place where brain scientists could find all relevant analyses and data, run each analysis on each dataset, and see a leaderboard comparing performances, without writing any lines of code. Cloud-based solutions simplify reproducibility and extensibility by essentially eliminating activation energy and extraneous sources of analytic variability.

The second goal is to enable such a system to work with "big data" (i.e., data too large to fit on a workstation). Data are scaling in many domains in brain science, either because individual experiments are large (as in calcium imaging and whole-brain CLARITY imaging), there are thousands of subjects with gigabytes of data each (as in large-scale human brain imaging projects), or there are millions of time points (as in wearable sensor data). Regardless of source and modality, if it is "medium data" (meaning too large to fit in memory, but small enough to fit on your computer), tasks as simple as visualizing, rotating, and opening the data are challenging using standard tools such as MATLAB, Python, or ImageJ. For big data, the challenges are even larger because questions of how to store, compress, manage, and archive the data exceed the computational capabilities and resources of most experimental labs. Cloud-based solutions simplify big data analysis due to their inherently scalable nature.

What's the Big Idea?

We are proposing to design, build, and deploy an instance of "cloud neuroscience," meaning that the data, the code, and the analytic results all live in the cloud together. Cloud neuroscience can be thought of as an operating system, a set of programs that run on it, a file system that stores the data, and the data itself, all designed to run in a scalable fashion and to be accessible from anywhere.

What Are the Design Criteria?

First and foremost, the design and construction should be organic, grassroots, and open source, to ensure that it remains intimately connected to the needs of all scientific citizens. Over 100,000 people attend annual brain science conferences, including neuroscience, psychology, psychiatry, and neurology. This is a massive human capital resource, so the system should enable contributions from any of them, regardless of background or resources. Thus, the system needs to support data and workflows of all kinds, regardless of modality, complexity, or scale—including raw data, derived data, and metadata. Doing so would also further democratize brain sciences, opening the door to the additional 3.5 billion people with mobile broadband access who could contribute if given the opportunity. Encouraging and supporting such

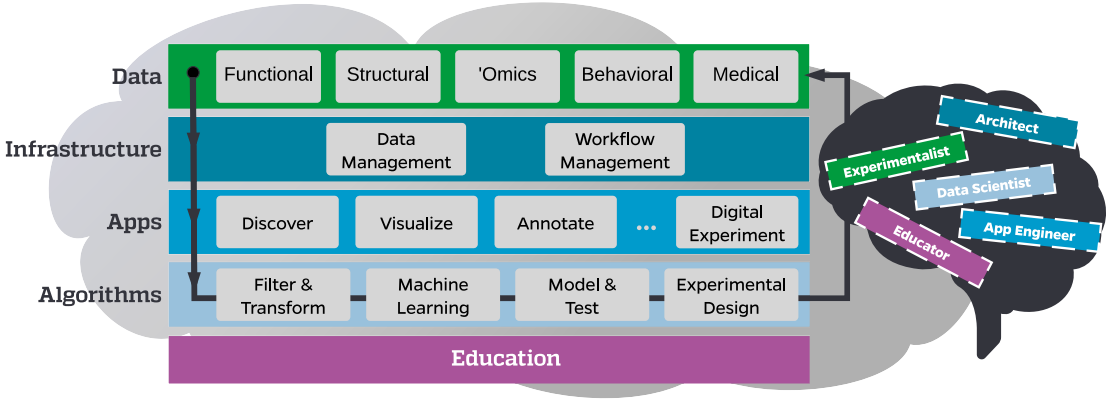

Figure 2. Schematic of the Five Proposed Components
An individual can adopt any or all of the five roles (color-coded dashed rectangles). For each component, the cloud content is generated by individuals in one of the five roles.

involvement motivates an emphasis on ethical standards and cultural sensitivities. Moreover, millions of hours and billions of dollars have been spent developing brain science resources, including vast quantities of data, algorithms, and models. The system should build upon such work. Because different people have different preferences, access controls should be flexible enough to satisfy everyone's needs. For resources that are open, reproducing and extending prior work should be "turn-key," allowing researchers to "swap in" different datasets or algorithms as desired. Industry is making tremendous headway in this regard, including digital notebooks to keep track of all analyses, software containers to ease the burden of installing and configuring software, and web services that dynamically provide computational resources as needed. To the extent possible, we should leverage these resources and engage with nonprofit, institutional, and corporate partners to express our domain-specific needs. The design should be highly adaptive, to capitalize on rapid advances from within and outside brain sciences, and, of course, open source with permissive licenses. And the entire system should be able to run not just in a single commercial cloud, but also on other clouds, national resources, institutional clusters, local workstations, and laptops, to enable maximal portability and utility. Perhaps most importantly, the system should be universally useful, helping to answer the grand challenges of brain science while facilitating much greater participation in the scientific process.

The motivation underlying this endeavor is to accelerate the scientific process by improving the experience of doing brain science. Thus, the community can determine the worst pain points in our process and design solutions around them. For example, if looking at data is the largest bottleneck, then one could use a cloud-based visualization app (like Google Maps, CATMAID, or NeuroDataViz). On the other hand, if the largest bottleneck is getting data into a common format before running analyses, then one would benefit from having all the data stored in a format with a standardized application programming interface (API) so every dataset can be accessed in the same way. In other words, it is time for the scientific community to prioritize the user experience to focus the subsequent software development.

How Might We Achieve It?

In this section, we propose a potential design of the constituent components that could comprise an instance of cloud neuroscience (see Figure 2). The required elements can be divided into five categories: data, infrastructure, apps, algorithms, and education. The goal of breaking down the problem this way is to ensure that all brain scientists, professional and citizen alike, can contribute to and benefit from the system. Crucial to success will be tight integration across components, each of which is described in some detail below. Some brain scientists are able to span the full range from design to analysis, including running experiments, analyzing data, making discoveries, and even writing articles. Such polymaths can seamlessly alternate between different roles. Others might be highly skilled in software engineering, but not data collection. To ensure that all brain scientists can contribute to this effort, we have organized types of activities according to the "role" of the individual performing those activities. These roles are not meant to be prescriptive; rather, they serve to help guide scientists to the different kinds of contributions they could make (see Box 1 for detailed description of the roles).

Data

The data component is intended to mitigate difficulties with storing and accessing data, regardless of the modality, scale, or complexity of the data. Anybody would be able to upload raw data, derived data, and metadata as they flow off the sensors and dynamically control access. Functionality would build on and incorporate existing brain science data repositories (Ascoli et al., 2007; Burns et al., 2013; Crawford et al., 2016; Poldrack et al., 2013; Teeters et al., 2008), as well as more general services (e.g., FigShare). Therefore, the technical challenges for small and large data storage and access, for the most part, already have reasonable solutions for many data types. The remaining challenges are to further lower the barrier to entry, making data upload and access easier, especially for multiterabyte datasets. Data contributions will be able to come from anyone and could be stored in a variety of accessible places to minimize transfer cost and time. Access controls would enable scalable sharing

Box 1. Roles

We enumerate six different roles for participants. Note that these are not characterizing individuals but roles that any individual can play. Roles differ in their degree of interest and expertise in various aspects of the scientific process, all of which are important.

- Experimentalist: A person in this role is acquiring data. This includes activities such as recruiting subjects and specifying inclusion guidelines (for human studies), experimental setup, subject care, and data acquisition, as well as some aspects of data management and quality control. In this role, a person has extensive knowledge of the experiment details, though computational acumen can be quite modest.
- Architect: A person in this role is developing the infrastructure component. In this role, professional software engineer skills are required. Architects work collaboratively on open-source repositories, possibly co-localized.
- App Engineer: A person in this role is writing apps. These apps might wrap algorithms written by the engineer or others. In this role, best practices of software development for science, including proper scientific documentation, are crucial.
- Data Scientist: A person in this role is writing and running algorithms. These algorithms might serve any step of the scientific process. Data scientists have a wide variety of computational backgrounds, including engineering, physics, mathematics, statistics, and computer science.
- Scientific User: A person in this role is using tools to analyze and understand the data. This can take many forms, ranging from looking at images and figures generated directly from the data acquisition system to fitting statistical models and combining multiple disparate datasets. In this role, computational acumen is not required. Familiarity with the data, experimental details, etc. can vary widely.
- Educator: A person in this role is either creating or presenting educational content, including documentation, tutorials, and massive online open courses, as well as running workshops, hackathons, and summer courses.

with minimal effort. Storage costs would be the responsibility of the data provider if the data are private; if public, others could financially contribute. In either case, economies of scale would reduce storage costs, and we would work with commercial clouds and national infrastructures to offset costs to the extent possible. The data storage formats would allow visualization and analysis at scale.

Data contribution would be desirable and possible from any lab, regardless of its financial resources or location. For example, some methods are relatively inexpensive, such as EEG, fNIRS, and wearable technologies. Moreover, certain important subpopulations are better represented in less wealthy countries, enabling unique contributions from those places. If the same measures are included in more expensive projects, analysis bridges could be established between the datasets. This would enhance translational research at a global scale. These factors would lead to important collaborations in which less wealthy countries could influence the content and usefulness of this effort (Neuroinformatics Collaboratory, 2016).

Data types would include raw, derived, and metadata (see Box 2 for additional details). Raw data include data from any kind of experiment, including functional, structural, omics (e.g., genetic and epigenetic), behavioral, and medical data. Every experiment will be given a unique data identifier. Medical data will be given special attention to ensure compliance with national guidelines for patient privacy. Each data type will yield a wide diversity of derived data, including summary statistics, matrices, networks, shapes, and more. Associated with each entry is a collection of metadata, including a community-driven controlled vocabulary, as well as custom ad hoc fields. Metadata on the derived data will include detailed provenance history. The system would be seeded with existing reference datasets spanning spatial, temporal, and phylogenetic scales, including data from the Human Brain Project, the Human Connectome Project, the Allen Institute for Brain Science's data portal, IARPA's MICrONs program, and more.

Infrastructure

The infrastructure component is intended to mitigate difficulties in finding data or tools, linking them together, installing software, managing computers, and reproducing and extending results. When the infrastructure is operational, much of the scientific process can be conducted from a tablet or smartphone, replacing the need to buy and maintain high-power computers or keep software up to date. The infrastructure is essentially the operating system upon which all the services would run, akin to NeuroDebian (Halchenko and Hanke, 2012), but designed specifically for the cloud. This virtual operating system will run in the commercial cloud, on institutional resources, national centers, or local workstations, regardless of hardware configuration (e.g., Mac, Windows, Linux, etc.). The software could be designed and written by a small and distributed team of architects to facilitate design decisions considering diverse use cases.

The infrastructure could be composed of two core sub-components. First, a data management system would store and organize all the data. This could include managing access, assigning digital object identifiers (DOIs), and supporting common data formats, and would be easily extensible to new or custom formats. Data could also be compressed with or without loss, as desired by the contributor. Technically, data would be stored in a set of databases optimized for different brain science use cases. Second, a workflow management system would store and organize analyses, leveraging existing web services such as Github and continuous integration to the extent possible. This would enable "digital experiments," including all stages of data processing. Crucially, such experiments could be done on different hardware platforms, applied to different data (by merely swapping the DOI), or use different algorithms (a similarly simple modification). All infrastructure services would have easy-to-use APIs to maximize utility and extensibility.

Box 2. Types of Brain Science Data

- Functional data are fundamentally temporal and dynamic. Whether univariate or multivariate, the standard operations to apply include zooming in time, subsampling, smoothing, and converting to other domains such as Fourier. Functional data also have a spatial domain, which links them to structural data. The subdivision between functional and structural data may be, for some data, ambiguous.
- Structural data are fundamentally spatial in nature, include 2D images, 3D volumes, and 4D and 5D hypervolumes for multispectral and/or time-varying data (spatiotemporal data, such as fMRI and calcium imaging, are both structural and functional). This can include structural images, as well as sparse fluorescent images, gene expression maps, etc. Standard operations for these data include compression, downloads of volumes of arbitrary sizes and shapes, maximum projections, averages, and more.
- Omics data are sequential and categorical, including the genome, epigenome, metabolome, and microbiome. Standard queries for genetic data include sequence compression, alignment, and comparisons. Omics data may also have a spatial domain (e.g., gene expression data).
- Behavioral data can be of several different types. For example, behavior can be captured via video capture (e.g., behavioral observation of children during play), time series of task events during physiological measurements, questionnaires (e.g., symptom checklists), performance testing instruments (e.g., the NIH Toolbox), and other devices (e.g., actigraphy and voice recorders). Each datum has unique qualities and, therefore, functionality.
- Medical data include all electronic health data, including semi-structured text. They are among the most challenging of data types to aggregate, for until recently, the vast majority of the field has relied on paper charts or poorly structured electronic health record (EHR) systems. Fortunately, regulatory and funding agencies are incentivizing the widespread use of EHRs, as well as common data elements that are more amenable to data aggregation for the purposes of discovery science (e.g., the eMerge Network). Additionally, informatics frameworks are being developed to safely link disparate EHR data (e.g., https://www.i2b2.org/), and calls for the creation of open APIs are gaining attention.

Apps

The apps component is intended to mitigate difficulties in maintaining software versions, paying for software, and finding tools appropriate to run on data. Apps are the programs that run on the system, akin to tools like Dropbox (to upload/download), Google Maps (to visualize), PubMed Central (to search for information), BLAST (to compare your data with other data), and pipelines (to process your data). Apps can be developed by anybody with minimal programming skills, due to the careful design of the APIs in the infrastructure. A specification would be formalized and quality standards agreed upon by the community of users to publish apps in the open app marketplace. Different apps would be designed for users with different backgrounds, roles, and goals. For example, apps targeted at people in the experimentalist role could include features to enable uploading, downloading, and managing access without having to learn the APIs. On the other hand, apps targeted at people in the data analysis role could include preprocessing data, fitting models, testing hypotheses, plotting results, and running digital experiments. General purpose apps would include tools to visualize, manipulate, and manually annotate data. These general purpose apps enable a much broader community of users to participate in the scientific process, including those without extensive technical training or financial resources.

Algorithms

The algorithms component is intended to mitigate difficulties in analyzing data with increasing scale or complexity. Recent advances in artificial intelligence, including distributed machine learning libraries and deep learning, could be leveraged here. Algorithms operate on simulated, measured, or derived data to produce transformed representations or summary statistics of the data. Algorithms can be written by anybody with minimal data-science skills, including many current brain scientists, without knowledge of this proposed system (unlike apps). Algorithms are essentially "wrapped" in apps to run and therefore inherit many of the conveniences of the system. We partition algorithms into three different types. Scalable data-processing algorithms can be applied to a wide variety of data types. These will be easily daisy-chained together to obtain pipelines, which can similarly be adapted to apply different algorithms or data. Because algorithms will be applied more generally to less familiar data, or less familiar algorithms will be applied to familiar data, quality assessment will be particularly important. This would include both qualitative dashboards providing figures and quantitative metrics to evaluate and compare performances along different metrics. Finally, to optimize resources and avoid duplicating efforts across labs, experiments will need to be useful for a large number of people. Experimental design will therefore be a key algorithmic component as well.

Education

Just like there is a learning curve when switching from Windows to Mac, so too switching from current practices to this system will involve a learning curve. Therefore, the success of this endeavor will depend on extensive educational material, including documentation, tutorials, online courses, hackathons, workshops, and summer courses. All the content will be designed to complement existing educational resources, such as Coursera courses. The variety of educational resources would reflect the backgrounds and skills of the user and contributor communities, with the goal of universal access. Because of this variety, community-driven cultural sensitivity guidelines would be posted for all contribution types.

Discussion

Here we describe an immediately actionable grassroots proposal to marry recent advances in neurodata acquisition with scalable cloud computing to accelerate the process of discovery by scientists independently of how well resourced they are (we have developed a proof-of-concept example using multimodal MRI data; see http://neurodata.io for details). There are several mechanisms by which Cloud Neuroscience may yield benefits. Global collaborations may become much simpler and therefore more prevalent. Open science may be facilitated, and the barriers and benefits to conducting open science may become more transparent by virtue of the design. Many models can be tested on the same dataset, and individual models can be subjected to greater diversity of data-based reality checks. In the near term, any effort that generates reference data of interest to a large segment of the community can benefit from Cloud Neuroscience. One example is the upcoming ~10 petabytes from the IARPA MICrONS program.

Several potential criticisms are worth addressing, and many details need to be fleshed out. Privacy concerns for human data will require careful additional thinking so that best practices of anonymization and security can be implemented—precedent is provided by ongoing large research initiatives (e.g., Jack et al., 2008; Murphy et al., 2010; Sarwate et al., 2014). A viable financial model will be required. Potential partners include national laboratories that could contribute computing and storage resources, or companies interested in providing cloud-based web services for specific scientific subdomains. Return on investment must be considered. Cosmology, molecular genetics, and plant biology (see http://www.cyverse.org/) are existing proofs that when designed well, such resources can a yield dramatic and positive impact on the field. Other cloud-computing neuroscience efforts that focus on the human brain are already underway, such as CBRAIN (Das et al., 2016) and the Human Brain Project. Such efforts are important; the proposed project has been designed to leverage the developments from those projects and extend them to address a greater diversity of brain science questions, species, data modalities, and functionalities.

The above plans and challenges suggest immediately actionable next steps. A field engineer has been appointed to develop a survey to determine which existing resources are most useful (pooling information from places like https://github.com/ and https://www.nitrc.org/) and what new resources would be most useful. A software engineer has agreed to contribute significant effort toward building a "Neuroscience as a Service" framework (the virtual operating system and apps described above) based upon existing related services. They will begin formalizing minimal specifications for all resources. We have also obtained private seed funding to hire an additional senior software engineer. To gather community feedback, we will be monitoring https://neurostars.org/ for any posts that contain the tag "neurostorm." Next, sustainable governance, funding, and advisory models will be devised.

Pablo Picasso famously quipped, "Every child is an artist. The problem is how to remain an artist once we grow up." As the next generation of brain scientists grows up, we have an opportunity to provide them with a canvas on which they can craft ever more creative portraits of our minds. Cloud neuroscience is one step we can take in that direction.

SUPPLEMENTAL INFORMATION

Supplemental Information includes a complete author list with affiliations and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.10.033.

ABOUT THE AUTHORS

Joshua T. Vogelstein is a neurostatistician; an Assistant Professor of Biomedical Engineering at Johns Hopkins University (JHU); and a member of the Institute for Computational Medicine, Center for Imaging Science, and Kavli Neuroscience Discovery Institute (KNDI). Brett Mensh founded Optimize Science, a science consulting agency, and is Scientific Advisor at Janelia Research Campus. Drs. Vogelstein and Mensh co-organized the Global Brain Workshop, an event in April 2016 with Richard Huganir, Professor and Director of the Department of Neuroscience and Director of KNDI, JHU, and Michael I. Miller, Herschel and Ruth Seder Professor and University Gilman Scholar, Director of the Center for Imaging Science, and Co-director of KNDI, JHU. All the co-authors were invited to the Global Brain Workshop on the basis of their international leadership spanning different spatial, temporal, and phylogenetic scales. They each subsequently volunteered to continue discussing this content for the ensuing weeks and months.

REFERENCES

Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.-E., Bludau, S., Bazin, P.-L., Lewis, L.B., Oros-Peusquens, A.-M., et al. (2013). Science *340*, 1472–1475.

Ascoli, G.A., Donohue, D.E., and Halavi, M. (2007). J. Neurosci. *27*, 9247–9251.

Burns, R., Roncal, W.G., Kleissas, D., Lillaney, K., Manavalan, P., Perlman, E., Berger, D.R., Bock, D.D., Chung, K., Grosenick, L., et al. (2013). Sci Stat Database Manag. http://dx.doi.org/10.1145/2484838.2484870.

Crawford, K.L., Neu, S.C., and Toga, A.W. (2016). Neuroimage *124* (Pt B), 1080–1083.

Das, S., Glatard, T., MacIntyre, L.C., Madjar, C., Rogers, C., Rousseau, M.-E., Rioux, P., MacFarlane, D., Mohades, Z., Gnanasekaran, R., et al. (2016). Neuroimage *124* (Pt B), 1188–1195.

Denk, W., and Horstmann, H. (2004). PLoS Biol. *2*, e329.

Grillner, S., Ip, N., Koch, C., Koroshetz, W., Okano, H., Polachek, M., Poo, M.-M., and Sejnowski, T.J. (2016). Nat. Neurosci. *19*, 1118–1122.

Halchenko, Y.O., and Hanke, M. (2012). Front. Neuroinform. *6*, 22.

Jack, C.R., Jr., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al. (2008). J. Magn. Reson. Imaging *27*, 685–691.

Kent, S.M. (1994). Science with Astronomical Near-Infrared Sky Surveys, N. Epchtein, A. Omont, B. Burton, and P. Persi, eds. (Springer), pp. 27–30.

Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S., and Kohane, I. (2010). J. Am. Med. Inform. Assoc. *17*, 124–130.

Neuroinformatics Collaboratory (2016). Neuroinformatics Collaboratory, http://www.neuroinformatics-collaboratory.org.

Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., and Milham, M.P. (2013). Front. Neuroinform. *7*, 12.

Sarwate, A.D., Plis, S.M., Turner, J.A., Arbabshirani, M.R., and Calhoun, V.D. (2014). Front. Neuroinform. *8*, 35.

Sejnowski, T.J., Churchland, P.S., and Movshon, J.A. (2014). Nat. Neurosci. *17*, 1440–1441.

Teeters, J.L., Harris, K.D., Millman, K.J., Olshausen, B.A., and Sommer, F.T. (2008). Neuroinformatics *6*, 47–55.

The Economist (2016). The future of computing. The Economist, http://www.economist.com/news/leaders/21694528-era-predictable-improvement-computer-hardware-ending-what-comes-next-future.

Vogelstein, J.T., Amunts, K., Andreou, A., Angelaki, D., Ascoli, G., Bargmann, C., Burns, R., Cali, C., Chance, F., Chun, M., et al. (2016). arXiv, arXiv:1608.06548, https://arxiv.org/abs/1608.06548.

Weber, M., Mickoleit, M., and Huisken, J. (2014). Methods Cell Biol. *123*, 193–215.

Neuron, Volume *92*

Supplemental Information

To the Cloud! A Grassroots Proposal to Accelerate Brain Science Discovery

Neuro Cloud Consortium

Joshua T. Vogelstein,[1,33,34,35,36,*] Brett Mensh,[2,3,5] Michael Häusser,[4] Nelson Spruston,[5] Alan C. Evans,[6] Konrad Kording,[7] Katrin Amunts,[8,9,10] Christoph Ebell,[10] Jeff Muller,[10] Martin Telefont,[10] Sean Hill,[11] Sandhya P. Koushika,[12] Corrado Calì,[13] Pedro Antonio Valdés-Sosa,[14,15] Peter B. Littlewood,[16] Christof Koch,[17] Stephan Saalfeld,[5] Adam Kepecs,[18] Hanchuan Peng,[17] Yaroslav O. Halchenko,[19] Gregory Kiar,[1,33] Mu-Ming Poo,[20] Jean-Baptiste Poline,[21] Michael P. Milham,[22,23] Alyssa Picchini Schaffer,[24] Rafi Gidron,[25] Hideyuki Okano,[26,27] Vince D. Calhoun,[28,29] Miyoung Chun,[30] Dean M. Kleissas,[31] R. Jacob Vogelstein,[32] Eric Perlman,[33] Randal Burns,[34,35] Richard Huganir,[36,37] and Michael I. Miller[1,33,37]

[1]Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Optimize Science, Mill Valley, CA 94941, USA
[3]UCSF Kavli Institute for Fundamental Neuroscience, San Francisco, CA 94143, USA
[4]Wolfson Institute for Biomedical Research and Department of Neuroscience, Physiology, and Pharmacology, University College London, Gower Street, London WC1E 6BT, UK
[5]Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, VA 20147, USA
[6]Montreal Neurological Institute, McGill University, 3801 University Street, Montreal, QC H3A 2B4, Canada
[7]Departments of Physical Medicine and Rehabilitation, Physiology, Applied Mathematics, and Biomedical Engineering, Northwestern University, 345 East Superior Street, Chicago, IL 60611, USA
[8]Institute for Neuroscience and Medicine, INM-1, Forschungszentrum Jülich, 52428 Jülich, Germany
[9]Cécile and Oskar Vogt Institute of Brain Research, University Hospital Duesseldorf, University Duesseldorf, 40225 Düsseldorf, Germany
[10]Human Brain Project, EPFL, 1202 Geneva, Switzerland
[11]Blue Brain Project, EPFL, Campus Biotech, 1202 Geneva, Switzerland
[12]Department of Biological Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Navy Nagar, Colaba, Mumbai 400005, India
[13]Biological and Environmental Science and Engineering, KAUST, Thuwal 23955-6900, Saudi Arabia
[14]University of Electronic Science and Technology of China, Shahe Campus, Chengdu, Sichuan 610054, PRC
[15]Cuban Neurosciences Center, Cubanacan, Playa, Havana CP 11600, Cuba
[16]Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
[17]Allen Institute for Brain Science, 615 Westlake Avenue North, Seattle, WA 98109, USA
[18]Cold Spring Harbor Laboratory, Marks Building, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[19]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755, USA
[20]Institute of Neuroscience, Chinese Academy of Sciences Center for Excellence in Brain Science and Intelligence Technology, 320 Yue Yang Road, Shanghai 200031, China
[21]Henry H. Wheeler Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA 94720, USA
[22]Center for the Developing Brain, Child Mind Institute, 445 Park Avenue, New York, NY 10022, USA
[23]Nathan S. Kline Institute for Psychiatric Research, 140 Old Orangeburg Road, Orangeburg, NY 10962, USA
[24]Simons Collaboration on the Global Brain, Simons Foundation, 160 Fifth Avenue, 7th Floor, New York, NY 10010, USA
[25]Israel Brain Technologies, Precede Building, Hakfar Hayarok, Ramat Hasharon 47800, Israel
[26]Department of Physiology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan
[27]Laboratory for Marmoset Neural Architecture, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
[28]The Mind Research Network, Albuquerque, NM 87106, USA
[29]Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87131, USA
[30]The Kavli Foundation, 1801 Solar Drive, Suite #250, Oxnard, CA 93030, USA
[31]Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723, USA
[32]Intelligence Advanced Research Projects Activity (IARPA), Maryland Square Research Park, 5850 University Research Court, Riverdale Park, MD 20737, USA
[33]Center for Imaging Science
[34]Department of Computer Science
[35]Institute for Data Intensive Engineering and Science

Johns Hopkins University, Baltimore, MD 21218, USA
[36]Department of Neuroscience, Johns Hopkins University, Baltimore, MD 21205, USA

[37]Kavli Neuroscience Discovery Institute, Johns Hopkins University, Baltimore, MD 21218, USA

*Correspondence: jovo@jhu.edu