Similarity network fusion for aggregating data types on a genomic scale

Bo Wang[1,5], Aziz M Mezlini[1,2], Feyyaz Demir[1,2], Marc Fiume[2], Zhuowen Tu[3], Michael Brudno[1,2], Benjamin Haibe-Kains[4,5] & Anna Goldenberg[1,2]

Recent technologies have made it cost-effective to collect diverse types of genome-wide data. Computational methods are needed to combine these data to create a comprehensive view of a given disease or a biological process. Similarity network fusion (SNF) solves this problem by constructing networks of samples (e.g., patients) for each available data type and then efficiently fusing these into one network that represents the full spectrum of underlying data. For example, to create a comprehensive view of a disease given a cohort of patients, SNF computes and fuses patient similarity networks obtained from each of their data types separately, taking advantage of the complementarity in the data. We used SNF to combine mRNA expression, DNA methylation and microRNA (miRNA) expression data for five cancer data sets. SNF substantially outperforms single data type analysis and established integrative approaches when identifying cancer subtypes and is effective for predicting survival.

Rapidly evolving technologies are making it progressively easier to collect multiple and diverse genome-scale data sets to address clinical and biological questions. For example, large-scale efforts by The Cancer Genome Atlas (TCGA) have already amassed genome, transcriptome and epigenome information for over 20 cancers from thousands of patients. The availability of such a wealth of data makes integrative methods essential for capturing the heterogeneity of biological processes and phenotypes, leading to, for example, the identification of homogeneous subtypes in breast cancer. Data-integration methods need to overcome at least three computational challenges: (i) the small number of samples compared to the large number of measurements; (ii) the differences in scale, collection bias and noise in each data set, and (iii) the complementary nature of the information provided by different types of data. Current integration approaches have yet to address all of these challenges together[1–4].

The simplest way to combine biological data is to concatenate normalized measurements from various biological domains, such as mRNA expression and DNA methylation, for each sample. Unfortunately, concatenation further dilutes the already low signal-to-noise ratio in each data type. To avoid this, a common strategy is to analyze each data type independently[2,3,5,6] before combining data. However, such independent analyses often lead to inconsistent conclusions that are hard to integrate. Another approach to increase signal is to preselect a set of important genes from each data source and use Consensus Clustering[1] to combine the data[3]. However, preselecting genes leads to a biased analysis, and focusing only on common patterns can miss valuable complementary information. One recent machine-learning approach, iCluster[7], uses a joint latent variable model for integrative clustering. Though powerful, iCluster and related machine-learning approaches[4] do not scale to the full spectrum of available measurements, making the methods sensitive to the gene preselection step.

Our SNF approach is distinct in that it uses networks of samples as a basis for integration. For example, when combining data from patient samples, SNF creates a patient network. Although networks of individuals have been extensively studied in other contexts, most notably in social science[8] or in relation to disease[9], to our knowledge patient-similarity networks have not been used specifically for integrating biological data. SNF consists of two main steps: construction of a sample-similarity network for each data type and integration of these networks into a single similarity network using a nonlinear combination method.

The fused network captures both shared and complementary information from different data sources (**Supplementary Results** and **Supplementary Figs. 1–3**), offering insight into how informative each data type is to the observed similarity between samples. Because it is based on networks of samples, SNF can derive useful information even from a small number of samples, is robust to noise and data heterogeneity, and scales to a large number of genes. In addition to integrating data, our fused networks can efficiently identify subtypes among existing samples by clustering and predict labels for new samples based on the constructed network

[1]Genetics and Genome Biology, SickKids Research Institute, Toronto, Ontario, Canada. [2]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [3]Department of Cognitive Science, University of California San Diego, San Diego, California, USA. [4]Institut de Recherches Cliniques de Montréal, Université de Montréal, Montréal, Quebec, Canada. [5]Present addresses: Department of Computer Science, Stanford University, Stanford, California, USA (B.W.). and Ontario Cancer Institute, Princess Margaret Cancer Centre—University Health Network, Toronto, Ontario, Canada (B.H.-K.). Correspondence should be directed to A.G. (anna.goldenberg@utoronto.ca).

RECEIVED 8 MAY 2013; ACCEPTED 17 DECEMBER 2013; PUBLISHED ONLINE 26 JANUARY 2014; DOI:10.1038/NMETH.2810

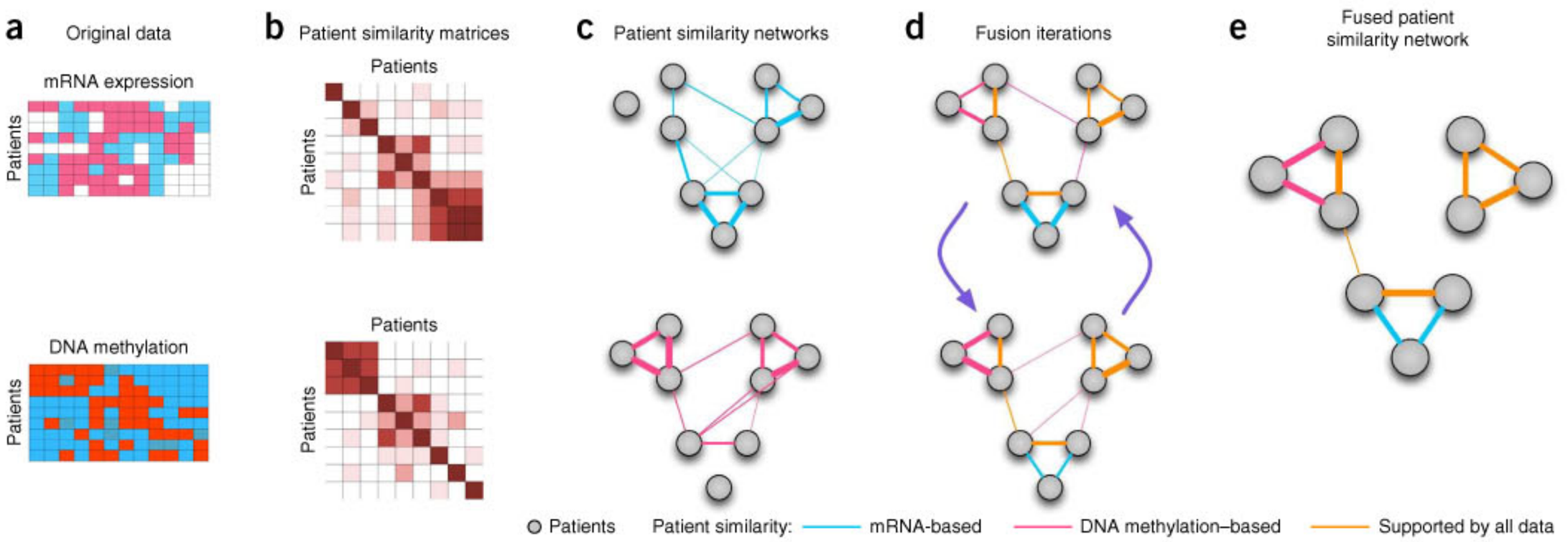


Figure 1 | Illustrative example of SNF steps. (**a**) Example representation of mRNA expression and DNA methylation data sets for the same cohort of patients. (**b**) Patient-by-patient similarity matrices for each data type. (**c**) Patient-by-patient similarity networks, equivalent to the patient-by-patient data. Patients are represented by nodes and patients' pairwise similarities are represented by edges. (**d**) Network fusion by SNF iteratively updates each of the networks with information from the other networks, making them more similar with each step. (**e**) The iterative network fusion results in convergence to the final fused network. Edge color indicates which data type has contributed to the given similarity.

(Online Methods, **Supplementary Note 1** and **Supplementary Fig. 4**). Combining diverse data types from five different human cancers, we demonstrated that SNF yields coherent, clinically relevant patient subtypes and improves on the performance of popular integrative approaches and a network-based approach that uses individual data types. The SNF software easily scales to multiple genome-wide data types with tens of thousands of measurements and is freely available as **Supplementary Software** and at http://compbio.cs.toronto.edu/SNF/.

RESULTS

Method overview

Given two or more types of data for the same set of samples (e.g., patients), SNF first creates a network for each data type and then fuses these into one similarity network. The initial step is to use a similarity measure for each pair of samples to construct a sample-by-sample similarity matrix for each available data type (**Fig. 1a**,**b**). The matrix is equivalent to a similarity network where nodes are samples (e.g., patients) and the weighted edges represent pairwise sample similarities (**Fig. 1c**). Both matrices and networks are effective visual representations: similarity matrices help identify global patterns (clusters), whereas networks emphasize the detailed similarity patterns and the types of data that support each edge.

The network-fusion step (**Fig. 1d**) uses a nonlinear method based on message-passing theory[10] that iteratively updates every network, making it more similar to the others with every iteration. After a few iterations, SNF converges to a single network (**Fig. 1e**). The empirical convergence for a variety of data sets is shown on **Supplementary Figures 5–7**. The method is robust to a variety of the hyperparameter settings (Online Methods and **Supplementary Figs. 8–10**). The advantage of our integrative procedure is that weak similarities (low-weight edges) disappear, helping to reduce the noise (**Fig. 2** and **Supplementary Fig. 2**), and strong similarities (high-weight edges) present in one or more networks are added to the others. Additionally, low-weight edges supported by all networks are retained depending on how tightly connected their neighborhoods are across networks. Such nonlinearity allows SNF to make full use of a network's local structure, integrating common as well as complementary information across networks.

A case study: glioblastoma multiforme

Multiple integrative approaches have been applied to understand the heterogeneity and identify the subtypes of glioblastoma multiforme (GBM), an aggressive adult brain tumor. Depending on the type of data used, these integrative analyses often lead to different conclusions. For example, one analysis that had combined expression and copy-number-variant data had identified two subtypes[11], but a later analysis[2], driven primarily by expression data, had identified four subtypes, which does not agree with the previous findings. A recent DNA methylation–based approach had identified three subtypes: one characterized by a somatic mutation in *IDH1* (ref. 12) and two others roughly corresponding to the subtypes identified in ref. 2. Though methylation data had been used for the analysis in ref. 2, the IDH subtype had not been identified because of the expression data–driven subtyping analysis.

We used SNF to fuse three data types for 215 patients with GBM: DNA methylation (1,491 genes), mRNA expression (12,042 genes) and miRNA expression (534 miRNAs). As expected, networks built using a single data type yielded very different patterns supports of patient similarity. For example, DNA methylation strongly supports connectivity in the smallest patient cluster (**Fig. 2a**), whereas mRNA expression supports similarity in the medium-sized cluster (**Fig. 2b**). DNA methylation and mRNA expression suggest relatively strong intercluster similarity (**Fig. 2a**,**b**), though the exact patterns are different between those data types. It is difficult to discern patterns in the patient-similarity network based on miRNA data alone (**Fig. 2c**). The fused network gives a much clearer picture of clustering in our set of patients with GBM, illustrated by the tightness of connectivity within clusters and relatively few edges between clusters (**Fig. 2d**).

We unified the results of several previous GBM analyses as well as identified new and potentially interesting associations. For example, our smallest cluster (subtype 3) corresponds to the previously identified IDH subtype[12] consisting of younger patients with a substantially more favorable prognosis. All patients with

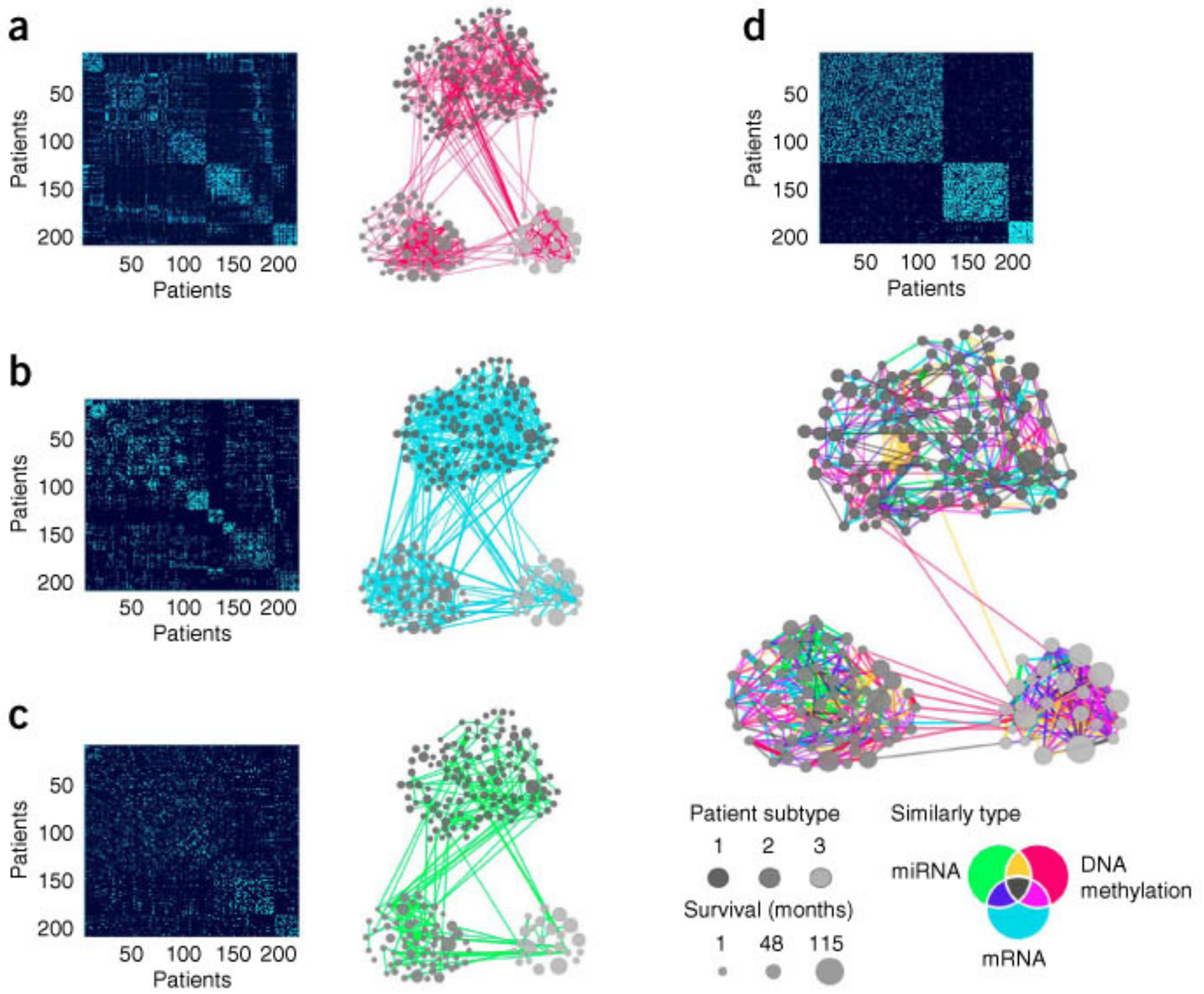


Figure 2 | Patient similarities for each of the data types independently compared to SNF fused similarity. (**a**–**d**) Patient-to-patient similarities for 215 patients with GBM represented by similarity matrices and patient networks, where nodes represent patients, edge thickness reflects the strength of the similarity, and node size represents survival. Clusters are coded in grayscale (subtypes 1–3) and arranged according to the subtypes revealed through spectral clustering of the combined patient network. The clustering representation is preserved for all four networks to facilitate visual comparison. DNA methylation (**a**), mRNA expression (**b**), miRNA expression (**c**) and SNF-combined similarity matrix and network (**d**; see **Supplementary Fig. 11** for more information about network edges).

an *IDH1* mutation for whom the information was available ($n = 14$ patients, Fisher exact test $P = 4.87 \times 10^{-11}$) belong to this cluster. Subtype 1 patients (hazards ratio (HR) = 0.278, Cox log-rank test $P = 0.001$; **Supplementary Fig. 11c**) had a favorable response to temozolomide (TMZ), a drug commonly used to treat GBM (**Supplementary Figs. 11** and **12**). One of the reasons for the lack of such an effect in subtype 2 could be its significant association with *CTSD* overexpression ($P < 0.001$, Bonferroni-corrected), which has been found to prevent the effect of TMZ *in vitro*[13] (**Supplementary Results**).

Our network analysis goes beyond subtyping. Each edge in the fused network is colored by the data type(s) that contributed to the given similarity. A multicolor cluster means that no single data type or combination support patient similarity across GBM. We found that most edges were supported by at least two data types: 49.5% of all patient similarities (edges) were due to two data types, 17.2% were supported by all three data types and the remaining 33.3% of the edges were supported by only one data type, with strong enough similarity that those edges remained prominent in the fused network (**Supplementary Fig. 13**). The GBM analysis highlights three important features of our network-based integrative approach: (i) the ability to detect common as well as complementary signals (**Fig. 2d** and **Supplementary Fig. 1**); (ii) the ability to reduce noise by aggregating across multiple types of data (**Fig. 2d**, and **Supplementary Figs. 2** and **3**); and (iii) insight into the relative importance of each data source for determining patient similarity, thus refining our understanding of the heterogeneity within each subtype (**Fig. 2d** and **Supplementary Fig. 14**).

Evaluating SNF across a wide spectrum of cancers

In addition to the GBM analysis, we applied SNF to four other cancer profiles by TCGA: breast invasive carcinoma (BIC), kidney renal clear cell carcinoma (KRCCC), lung squamous cell carcinoma (LSCC) and colon adenocarcinoma (COAD). The DNA methylation, mRNA and miRNA expression data for these cancers vary in sample size (from 92 for COAD to 215 for GBM) and number of measurements (from 534 miRNAs in GBM to 27,578 methylated genes in LSCC and COAD) as well as heterogeneity[3,5,6] (**Supplementary Data** and **Supplementary Table 1**).

We evaluated SNF performance by identifying subtypes in each of these cancers. We report three commonly used measures: (i) P value in Cox log-rank test to evaluate the significance of the difference in survival profiles between subtypes[14]; (ii) silhouette score[15], a measure of cluster coherence, to evaluate whether patients are more similar within or across subtypes; and (iii) algorithm running time to evaluate scalability (**Supplementary Note 2**). We used spectral clustering (Online Methods) on the patient network to identify homogeneous cancer subtypes. We compared SNF to iCluster[7] and the concatenation of the three types of data (**Supplementary Note 3**).

We first compared data integration to the use of individual data types separately across the five cancers. We obtained patient clusters for individual data types by building a patient-similarity network and clustering it using spectral clustering (same as for SNF). Except for a few cases, single data type analysis did not lead to significantly different survival profiles, but networks fused by SNF had significant differences in survival among subtypes in all five cancers (**Table 1**). Note that the added fusion step is the only difference between the single and fused analyses. Spatial embedding of the subtypes from the fused network for each cancer showed very clear separation between clusters (**Supplementary Fig. 15**).

One major limitation of current integrative methods such as iCluster is the need for *a priori* gene selection. Although SNF

Table 1 | SNF-based analysis versus individual data types

Cancer type	mRNA expression	DNA methylation	miRNA	SNF
GBM (3 clusters)	0.54	0.11	0.21	2.0×10^{-4}
BIC (5 clusters)	0.03	0.05	0.30	1.1×10^{-3}
KRCCC (3 clusters)	0.20	0.61	0.17	2.9×10^{-2}
LSCC (4 clusters)	0.06	0.26	0.46	2.0×10^{-2}
COAD (3 clusters)	0.18	0.04	0.46	8.8×10^{-4}

Analysis using Cox log-rank test P values.

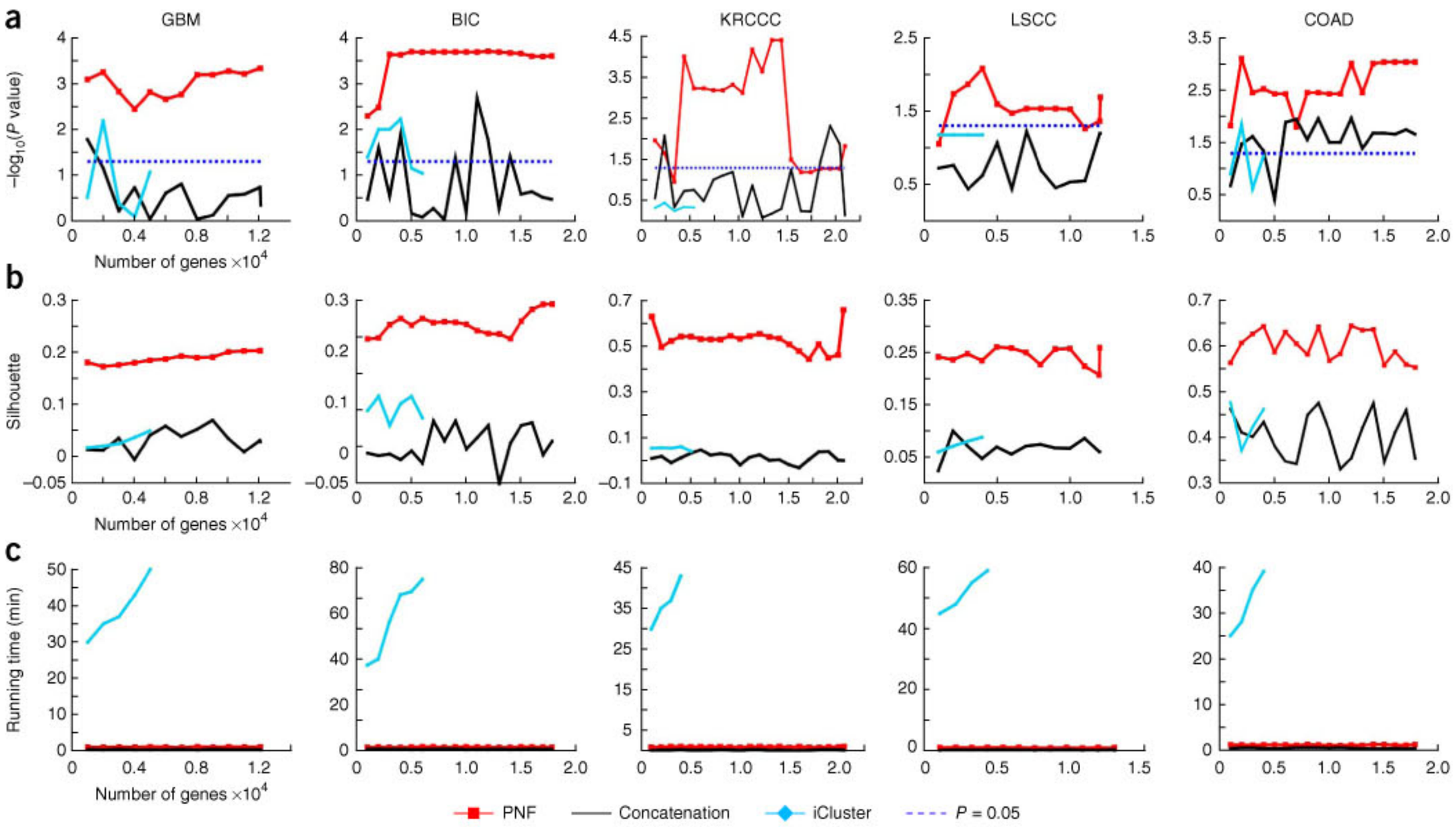


Figure 3 | Comparison of the SNF approach to iCluster and concatenation. (**a**–**c**) Cox log-rank test *P* value for survival analysis (**a**), silhouette score representing the coherence of clusters (**b**) and running time (**c**) for each of the indicated cancers. Number of preselected genes is shown on *x* axes.

does not require preselection, for comparison we report the performance of all three methods as a function of the number of preselected genes, ordering genes by significance for differential expression between tumor and healthy tissue using the significance analysis of microarrays (SAM) test[16] (**Supplementary Note 2**). SNF achieved significance in survival analysis across the spectrum of preselected genes (**Fig. 3a**) and resulted in substantially more coherent clusters according to the silhouette score (**Fig. 3b**). Comparative performance across cancers showed that in GBM and BIC, Cox survival *P* values were very stable with respect to the number of preselected genes. There is more fluctuation in survival *P* values for KRCCC and LSCC. This is explained by the fact that both KRCCC and LSCC have at least one subtype with very few patients (**Supplementary Fig. 15**), making the *P* values very sensitive to any change in clustering. This is a common problem of rare disease subtypes; the silhouette score in this case is a better indicator of clustering stability.

iCluster achieved significance for a small number of genes but was very sensitive to gene preselection. The performance of concatenation was even less predictable, though it was substantially faster, as illustrated by the running-time analysis (**Fig. 3c**). The computational complexity of the concatenation approach was equivalent to the time needed to run hierarchical clustering. Running time for SNF was only marginally higher than for concatenation. iCluster performance scaled exponentially in the number of genes, which explained the necessity for gene preselection.

From subtype-based to network-based outcome prediction

We showed that clustering patient networks derived from multiple data types using SNF performs as well or better than the state-of-the-art subtyping methods applied to survival analysis (**Supplementary Figs. 16–19**). On the harder task of survival risk prediction, we also found that subtyping was inferior to a true network-based approach (**Supplementary Fig. 20**).

We used the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) breast cancer data set[17] to validate our network-based prediction. The METABRIC data set consists of a discovery cohort (997 patients) and validation cohort (995 patients). We performed a PAM50 analysis (a standard breast cancer signature), iCluster analysis (InterClust[17]) and SNF analysis with five clusters (chosen by our model selection criterion) and ten clusters (for comparative purposes) (**Table 2**). The published[17] significance value obtained using iCluster on the validation set is lower than both the iCluster-based discovery cohort *P* value and the validation *P* value obtained using SNF, suggesting the potential for overfitting by iCluster. The concordance index (CI) is a continuous and robust accuracy measure to assess the prognostic

Table 2 | METABRIC survival analysis and prediction

	PAM50 (5 clusters)	iCluster (10 clusters)	SNF (5 clusters)	SNF (10 clusters)	Network
P value discovery cohort	3.0×10^{-9}	1.2×10^{-14}	6.10×10^{-11}	3.31×10^{-12}	–
P value validation cohort	1.7×10^{-9}	2.9×10^{-11}	5.12×10^{-13}	7.86×10^{-12}	–
CI discovery cohort	0.560	0.621	0.638	0.638	0.720
CI validation cohort	0.551	0.605	0.633	0.633	0.706

Comparison of SNF and alternative methods on survival analysis (Cox log-rank test *P* value) and risk of death prediction in the METABRIC data (CI, concordance index).

value of risk prediction models (**Supplementary Note 2**). CI for SNF was higher (better) than CI for PAM50 and iCluster on both discovery and validation cohorts for both five and ten clusters (**Table 2**). The CI values were relatively similar for all compared methods, indicating that subtype-based analyses have certain limitations.

We developed a network-based prediction approach that takes advantage of the whole network of patients rather than just individual clusters. Specifically, our network-based approach uses the fused network to constrain the Cox regression model to predict similar survival values for biologically similar patients (Online Methods and **Supplementary Results**). The network-based approach resulted in over 10% improvement in CI without any parameter tuning (**Table 2**). This network-based CI prediction on the validation cohort ranks in the top 20 of 1,400 models designed specifically for this task[18]. As we used the same network to assess CI for subtyping survival analysis and network-based survival analysis, we attribute the improvement in our results to the incorporation of richer information contained in the network.

DISCUSSION

We propose the SNF to integrate data in the space of samples (e.g., patients) rather than measurements (e.g., genes). Using SNF we constructed patient networks and combined mRNA expression, DNA methylation and miRNA expression data to identify subtypes with differential survival profiles. SNF also has many other applications. In the clinical domain, patient networks allow integration of very different kinds of measurements, such as microbiome and metabolomics data, questionnaires and functional magnetic resonance imaging, together with genomic, clinical and demographic data, as long as the data can be used to identify similarity between patients (Online Methods). Although some of these data types have been combined previously, our method enables their combination into a single comprehensive network that yields precise manifolds of diseases.

SNF can help answer questions that require combining multiple types or sources of data for the same set of objects or subjects, not just humans. For example, combining transcriptomic, epigenetic and genetic data for different tomato strains helps to visualize how biological similarity relates to the phenotype of interest, such as tomato sweetness. SNF can also integrate various gene-interaction data such as physical interactions, coexpression and colocalization data. In another context, it can improve the reliability and remove the experimental bias in constructing gene coexpression networks by integrating tissue-specific gene-expression data from a variety of experiments.

One important advantage of our approach is that it goes beyond current subtyping strategies to capture continuous phenotypes. Our analysis of cancers shows that although there are broad categories of patients (subtypes), the reality is more complex. Capturing variability in similarity and underlying biology via similarity networks moves us closer to the clinic of the future[19]. We believe that our fused networks will ultimately pave the way to much more refined representation and understanding of diseases, phenotypes and other biological phenomena.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This study used data generated by TCGA and METABRIC; we thank TCGA, the Cancer Research UK and the British Columbia Cancer Agency Branch for sharing these invaluable data with the scientific community. We thank N. Jabado, M. Wilson and J. Rommens for feedback on the manuscript, and B. Sousa for help with the figures. This study was partially funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-068) to M.B.; A.G. is funded by the SickKids Research Institute. Z.T. was supported by NSF IIS-1360568.

AUTHOR CONTRIBUTIONS

B.W. and A.G. conceived of and designed the approach. B.W. performed the data analysis, implemented the method in Matlab and performed all computational experiments. A.M.M. performed data preparation. F.D. wrote the R code that is distributed with the paper. M.F. assisted with network visualization and analysis. Z.T. helped with method design and theoretical framework. B.H.-K. assisted in preparation and analysis of the METABRIC data. B.W., M.B. and A.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
2. Verhaak, R.G.W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).
3. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
4. Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z. & Wild, D.L. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**, 3290–3297 (2012).
5. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
6. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
7. Shen, R., Olshen, A.B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
8. Goldenberg, A., Zheng, A.X., Fienberg, S.E. & Airoldi, E.M. A survey of statistical network models. *Foundations and Trends in Machine Learning.* **2**, 129–233 (2010).
9. Barabási, A.-L. Network medicine -from obesity to the 'diseasome. *N. Engl. J. Med.* **357**, 404–407 (2007).
10. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, 1988).
11. Nigro, J.M. *et al.* Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Res.* **65**, 1678–1686 (2005).
12. Sturm, D. *et al.* Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22**, 425–437 (2012).
13. Sun, S. *et al.* Protein alterations associated with temozolomide resistance in subclones of human glioblastoma cell lines. *J. Neurooncol.* **107**, 89–100 (2012).
14. Hosmer Jr, D.W., Lemeshow, S. & May, S. *Applied Survival Analysis: Regression Modeling of Time to Event Data* (Wiley, 2011).
15. Rousseeuw, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
16. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
17. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
18. Margolin, A.A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181 (2013).
19. Friend, S.H. & Ideker, T. Point: Are we prepared for the future doctor visit? *Nat. Biotechnol.* **29**, 215–218 (2011).

Independent optical excitation of distinct neural populations

Nathan C Klapoetke[1–5], Yasunobu Murata[4,5], Sung Soo Kim[6], Stefan R Pulver[6], Amanda Birdsey-Benson[4,5], Yong Ku Cho[1–5], Tania K Morimoto[1–5], Amy S Chuong[1–5], Eric J Carpenter[7], Zhijian Tian[8], Jun Wang[8], Yinlong Xie[8], Zhixiang Yan[8], Yong Zhang[8], Brian Y Chow[9], Barbara Surek[10], Michael Melkonian[10], Vivek Jayaraman[6], Martha Constantine-Paton[4,5], Gane Ka-Shu Wong[7,8,11] & Edward S Boyden[1–5]

Optogenetic tools enable examination of how specific cell types contribute to brain circuit functions. A long-standing question is whether it is possible to independently activate two distinct neural populations in mammalian brain tissue. Such a capability would enable the study of how different synapses or pathways interact to encode information in the brain. Here we describe two channelrhodopsins, Chronos and Chrimson, discovered through sequencing and physiological characterization of opsins from over 100 species of alga. Chrimson's excitation spectrum is red shifted by 45 nm relative to previous channelrhodopsins and can enable experiments in which red light is preferred. We show minimal visual system–mediated behavioral interference when using Chrimson in neurobehavioral studies in *Drosophila melanogaster*. Chronos has faster kinetics than previous channelrhodopsins yet is effectively more light sensitive. Together these two reagents enable two-color activation of neural spiking and downstream synaptic transmission in independent neural populations without detectable cross-talk in mouse brain slice.

Microbial opsins, which are light-driven ion pumps and light-gated ion channels that can be genetically expressed in excitable cells, enable the optical activation or inhibition of the electrical activity of defined neuron types, axonal pathways or brain regions with millisecond resolution[1–6]. There exists an unmet need to independently drive pairs of light-driven ion channels with different colors of light, which would enable the independent activation of different cell populations. Many groups have used channelrhodopsins with peak wavelength sensitivity to green light in conjunction with channelrhodopsins with peak wavelength sensitivity to blue light to achieve differential spiking with yellow light[7–10], but residual neural spiking cross-talk in response to blue light stimulation remains due to the intrinsic blue absorption by the retinal chromophore[11,12]. This fundamental limitation means it is currently not possible to achieve robust, temporally precise independent two-color spiking using a red-shifted channelrhodopsin alongside channelrhodopsin-2 (ChR2) in mammalian brain tissue. Previous efforts to further red-shift opsins and reduce blue light sensitivity through mutagenesis have empirically proven difficult, with spectra remaining little red shifted beyond that of the first reported green-peaked channelrhodopsin, VChR1 (refs. 9,13,14). Similarly, efforts to further increase the blue light sensitivity of ChR2 have resulted in much slower channelrhodopsins[15–17], which elicit low-temporal-precision spiking in response to light pulses of up to 1 s or longer.

One potential strategy for achieving independent two-color excitation is engineering differences in blue light sensitivity between blue and red light–drivable channelrhodopsins. Then a low blue irradiance could drive precisely timed spikes with the blue light–drivable channelrhodopsin while eliciting only subthreshold depolarizations (not causing synaptic release) in neurons expressing the red light–drivable channelrhodopsin. To achieve these goals, we turned to the natural world, performing *de novo* transcriptome sequencing of 127 species of alga.

Chronos is a new blue and green light–drivable channelrhodopsin with kinetics faster than those of previous channelrhodopsins as well as high light sensitivity. Chrimson is a new red light–drivable channelrhodopsin with spectra that are red shifted 45 nm more than those of previous channelrhodopsins. Together, Chronos and Chrimson allow independent two-color spike driving of, and synaptic release from, distinct neural populations in mouse brain slice. In addition, Chronos represents an excellent general-use channelrhodopsin. Chrimson may enable temporally precise experiments requiring red light, such as deep tissue

[1]The MIT Media Laboratory, Synthetic Neurobiology Group, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. [2]Department of Biological Engineering, MIT, Cambridge, Massachusetts, USA. [3]MIT Center for Neurobiological Engineering, MIT, Cambridge, Massachusetts, USA. [4]Department of Brain and Cognitive Sciences, MIT, Cambridge, Massachusetts, USA. [5]MIT McGovern Institute for Brain Research, MIT, Cambridge, Massachusetts, USA. [6]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia, USA. [7]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. [8]Beijing Genomics Institute-Shenzhen, Shenzhen, China. [9]Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [10]Institute of Botany, Cologne Biocenter, University of Cologne, Cologne, Germany. [11]Department of Medicine, University of Alberta, Edmonton, Alberta, Canada. Correspondence should be addressed to E.S.B. (esb@media.mit.edu) or G.K.-S.W. (gane@ualberta.ca).

RECEIVED 23 SEPTEMBER 2013; ACCEPTED 10 JANUARY 2014; PUBLISHED ONLINE 9 FEBRUARY 2014; DOI:10.1038/NMETH.2836

targeting, or scenarios in which blue light is visually distracting. To the latter end, we demonstrate brain stimulation through the cuticle of *Drosophila* and reduction in visual system–triggered responses during optogenetic experiments. These tools may also serve as protein backbones for building future optogenetic tools with novel capabilities.

RESULTS

Discovering novel channelrhodopsins via *de novo* sequencing

In recent years, a number of channelrhodopsins have been engineered for neuroscientific applications[18], derived from four channelrhodopsin genes from *Chlamydomonas reinhardtii* or *Volvox carteri*. However, all known natural channelrhodopsins have blue-green (430- to 550-nm) spectral peaks[13,14,18,19], and engineered red-shifted channelrhodopsins such as C1V1 (ref. 9) and ReaChR[14] have peak wavelength sensitivity to green light (~545 nm) and a similar spectrum to VChR1 (ref. 7). Furthermore, existing channelrhodopsins exhibit an inverse relationship between two desired properties: high light sensitivity and fast kinetics[18]. We therefore sought to overcome these limitations by exploring the genetic diversity of natural channelrhodopsins.

We *de novo*–sequenced 127 algal transcriptomes[20] and identified 61 channelrhodopsin homologs, which we subsequently synthesized and screened for photocurrents in HEK293 cells via whole-cell patch clamp (**Supplementary Figs. 1–4** and **Supplementary Table 1**). Of these, we selected opsins with novel characteristics for further characterization in cultured neurons (**Fig. 1** and **Supplementary Table 2**), focusing primarily on photocurrent, wavelength sensitivity, kinetics and trafficking (**Fig. 1** and **Supplementary Figs. 5–9**). To avoid selection bias, we cotransfected all opsins into neurons with a secondary tdTomato plasmid, and we selected cells solely according to the presence of cytosolic tdTomato expression (**Supplementary Fig. 5a**,**b**). This unbiased selection method was applied throughout the paper in all culture experiments unless otherwise indicated.

We assessed wavelength sensitivity and photocurrent amplitude, using ChR2 for blue (470 nm) comparison and $C1V1_{TT}$ (ref. 9) for green (530 nm) and far-red (660 nm) comparison (**Fig. 1a–d**). Of the 20 opsins screened in neurons, we found 4 previously unknown channelrhodopsins, one each from the species *Chloromonas oogama* (CoChR), *Chloromonas subdivisa* (CsChR), *Stigeoclonium helveticum* (ShChR) and *Scherffelia dubia* (SdChR) that bore either significantly higher blue photocurrents than ChR2 ($P < 0.001$; ANOVA with Dunnett's *post hoc* test used for all multiway comparisons; **Fig. 1c**) or significantly higher green photocurrents than $C1V1_{TT}$ ($P < 0.001$; **Fig. 1b**). Additionally, we discovered the first reported yellow-peaked channelrhodopsin, CnChR1 from the species *Chlamydomonas noctigama*. CnChR1 had 660-nm far-red light photocurrents of 674 ± 120 pA ($n = 11$ cells; values throughout are mean ± s.e.m.), which are significantly higher (~30×, $P < 0.0001$; **Fig. 1a**,**d**) than those of $C1V1_{TT}$. Owing to its spectral sensitivity, we nicknamed this molecule 'Chrimson'. With a spectral peak at 590 nm, Chrimson is 45 nm more red shifted than previously known channelrhodopsins (**Fig. 1e** and **Supplementary Figs. 5d**,**e** and **9**).

Kinetic parameters and spiking performance

The ability to optically evoke spikes necessitates that channelrhodopsins possess not only photocurrents sufficient to depolarize the neuron cell membrane above its spike threshold, but also on-, off- and recovery kinetics fast enough to precisely control spike timing and fidelity[18,21]. Previously published green and red light–drivable channelrhodopsins have relatively slow off-kinetics, which limits their utility for high-frequency neural activation[14,18]. We characterized the kinetic properties of opsins with green photocurrents comparable to or higher than those of $C1V1_{TT}$ and found that only CsChR and ShChR had faster turn-on, turn-off and recovery kinetics (**Fig. 1f–h**). With a turn-on of 2.3 ± 0.3 ms ($n = 8$ cells) and a turn-off of 3.6 ± 0.2 ms ($n = 7$ cells), the *S. helveticum* channelrhodopsin ShChR possesses the fastest reported kinetics to date (**Supplementary Fig. 3**): we therefore nicknamed this molecule 'Chronos'.

We assessed Chronos's green light (530-nm) spiking fidelities at various irradiances and frequencies in cultured neurons (**Fig. 2a–c** and **Supplementary Figs. 10** and **11**). As expected from its fast kinetic properties, Chronos-mediated optical spiking replicated electrically driven spiking between 5 and 60 Hz (**Supplementary Fig. 11**). In contrast, CsChR reliably drove spikes only up to 20 Hz, and $C1V1_{TT}$ could not reliably drive spikes above 10 Hz even at the highest expression level (**Supplementary Figs. 10** and **11**). It has previously been noted that slow off-kinetics or poor recovery kinetics of channelrhodopsins can cause depolarization block or reduce photocurrents over sustained pulse trains[18]. Consistent with this observation, the spike failures we observed at high frequencies for CsChR and $C1V1_{TT}$ primarily occurred later in the pulse train (**Supplementary Fig. 10e**).

We next examined red light (625-nm)-evoked spiking fidelities. Consistent with the earlier photocurrent screening, Chrimson was the only opsin screened capable of red light–driven spiking with 5-ms pulses (**Fig. 2d**): these irradiances and pulse widths resulted in Chronos depolarizations of less than 1.5 mV (**Supplementary Fig. 12g**). However, Chrimson's slow kinetics of current decay after the cessation of light, or τ_{off}, of 21.4 ± 1.1 ms ($n = 11$ cells), in conjunction with its poor recovery kinetics, caused depolarization block and channelrhodopsin inactivation at frequencies exceeding 10 Hz (**Fig. 2e** and **Supplementary Fig. 12a–d**). We therefore optimized these parameters via mutagenesis, and we identified the K176R mutant, denoted as ChrimsonR, which sped up the off-kinetics to 15.8 ± 0.4 ms ($n = 5$ cells) without altering the red-shifted action spectrum (**Fig. 2f** and **Supplementary Fig. 12e**,**f**). This kinetics improvement enabled fast, reliable red light–driven spiking at frequencies of at least 20 Hz in both cultured neurons and acute cortical slice (**Fig. 2e** and **Supplementary Fig. 13**), comparable to the blue light spiking performance of the commonly used ChR2(H134R)[18,22]. We additionally found ChrimsonR to be capable of reliably eliciting spikes in cortical slice using >20-ms pulses of far-red light (735 nm) (**Supplementary Fig. 13**), which may be useful for *in vivo* scenarios in which deep-tissue light penetration, or lack of visual drive, is desired by the experimenter.

Use of Chrimson in *Drosophila*

Although optogenetic tools have become widely used in mammalian behavioral experiments, such tools have found more limited use in *Drosophila* experiments[23–27], possibly owing to strong innate behavioral artifacts induced by stimulation light reaching photosensitive regions[28] and poor cuticle penetration of blue light[29]. Optogenetics has therefore typically been used (i) in constrained circumstances in which light is delivered to

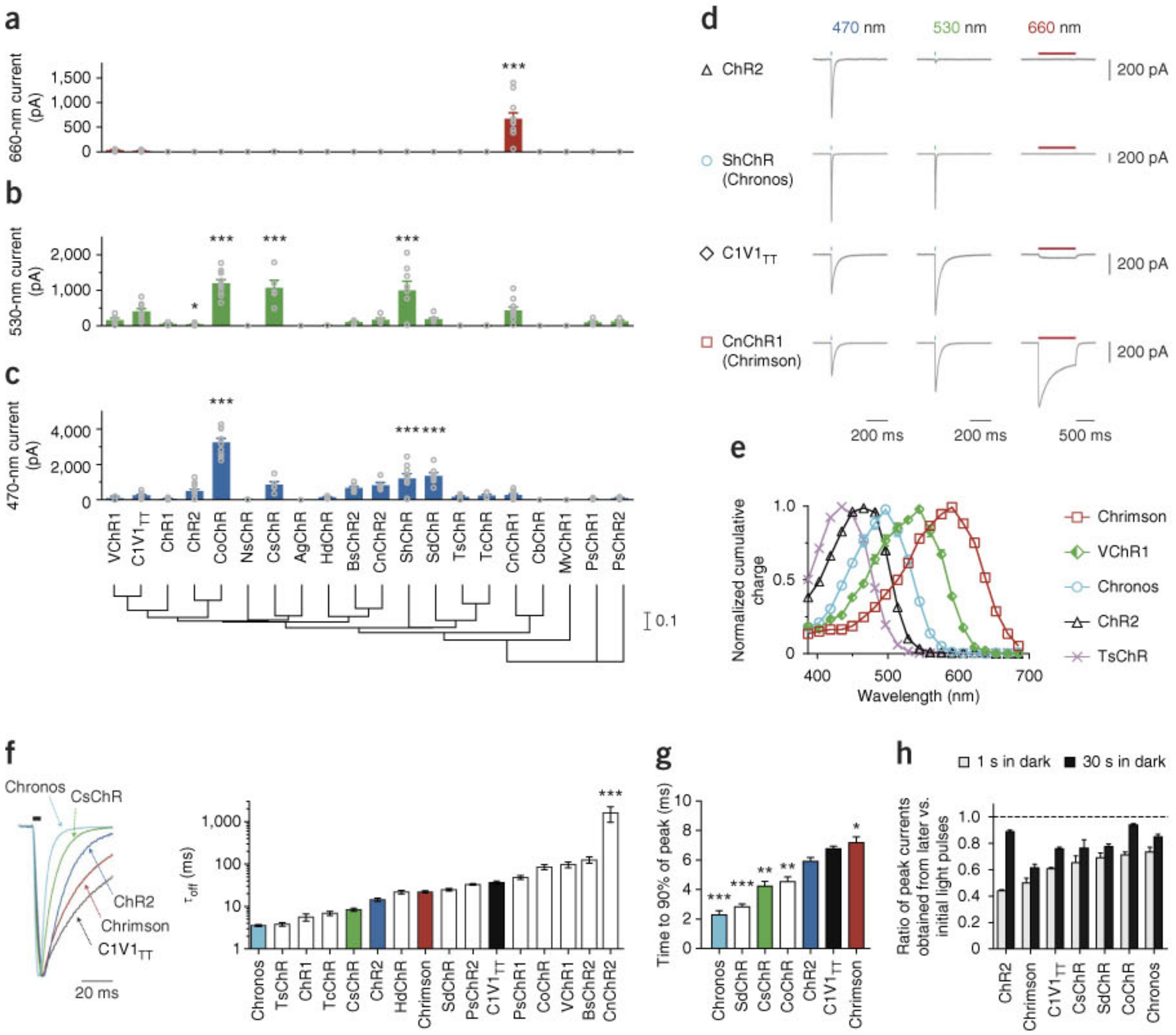


Figure 1 | Novel channelrhodopsin spectral classes discovered through algal transcriptome sequencing. (**a**–**c**) Maximum photocurrents in cultured neurons transfected with the different opsin-GFP fusions in response to far-red (660-nm), green (530-nm) and blue (470-nm) light; blue and green photon fluxes were matched, with illumination conditions defined as follows: 1-s pulse at 10 mW/mm^2 for red, 5-ms pulse at 3.66 mW/mm^2 for green and 5-ms pulse at 4.23 mW/mm^2 for blue. Individual cell photocurrents are plotted as gray circles and overlaid on the population bar graph. See **Supplementary Figure 6** for additional individual cell data. Bottom, phylogeny tree of the channelrhodopsins tested, based on transmembrane helix alignments. The scale indicates the number of substitutions per site. (**d**) Representative voltage-clamp traces in cultured neurons as measured under the screening conditions in **a**–**c** (the long red light pulse was used to ensure we did not miss any red-sensitive channelrhodopsins in our screen). (**e**) Channelrhodopsin action spectra (HEK293 cells; n = 6–8 cells; measured using equal photon fluxes, ~2.5 × 10^{21} photons/s/m^2). (**f**–**h**) Channelrhodopsin kinetic properties as measured in cultured neurons (see also **Supplementary Figs. 7** and **8**). Off-kinetics (**f**) were measured under the conditions in **a**–**c**; on-kinetics (**g**) and recovery kinetics (**h**) were measured with a 1-s pulse at 5 mW/mm^2. All opsins were illuminated near their respective peak wavelength, which was either blue or green for all opsins except Chrimson, which was characterized at 625 nm (n = 5–12 cells for all kinetic comparisons). τ_{off}, monoexponential fit of photocurrent decay. Peak current recovery ratios in **h** were determined from three 1-s light pulses, with the first pulse response used as the baseline for peak current recovery ratio calculations for both the second (1 s in dark after first pulse) and third pulse response (30 s in dark after second pulse). *P < 0.05, **P < 0.01, ***P < 0.001; ANOVA with Dunnett's *post hoc* test, with ChR2 as the reference in **c**,**f**,**g**, and C1V1$_{TT}$ as the reference in **a**,**b**. Exact P values and n values are in **Supplementary Table 2**. Plotted data are mean ± s.e.m. Opsins and the species they derive from are defined in the Online Methods.

peripheral organs while being blocked from reaching the eyes[25], (ii) in blind flies[23] or (iii) for cases in which undesired side effects of visible light stimulation do not substantially affect experimental interpretation[30]. Other technologies such as thermogenetics are often used despite their much slower time course[31,32]. We conjectured that red light activation with Chrimson, which is ~45 nm more red shifted than ReaChR[14,29] (**Supplementary Fig. 9**), would extend the reach of optogenetic tools to *Drosophila* behavioral experiments with reduced light-induced behavioral artifact and improved cuticle penetration.

We first used the *Drosophila* larval neuromuscular junction (NMJ) to examine the reliability of light-triggered action potentials in fly axons expressing Chrimson. Larval muscles have passive membrane properties, so excitatory junction potentials (EJPs) at the larval NMJ accurately reflect spiking in motor axons[31]. In Chrimson-expressing larvae, both 470-nm and 617-nm light triggered EJPs, even at short (1- to 2-ms) light-pulse durations and low intensities (0.06–0.14 mW/mm^2). In response to light-pulse durations over 2 ms, Chrimson activation triggered long-lasting barrages of EJPs (**Fig. 3a**,**b**). Long-wavelength red light

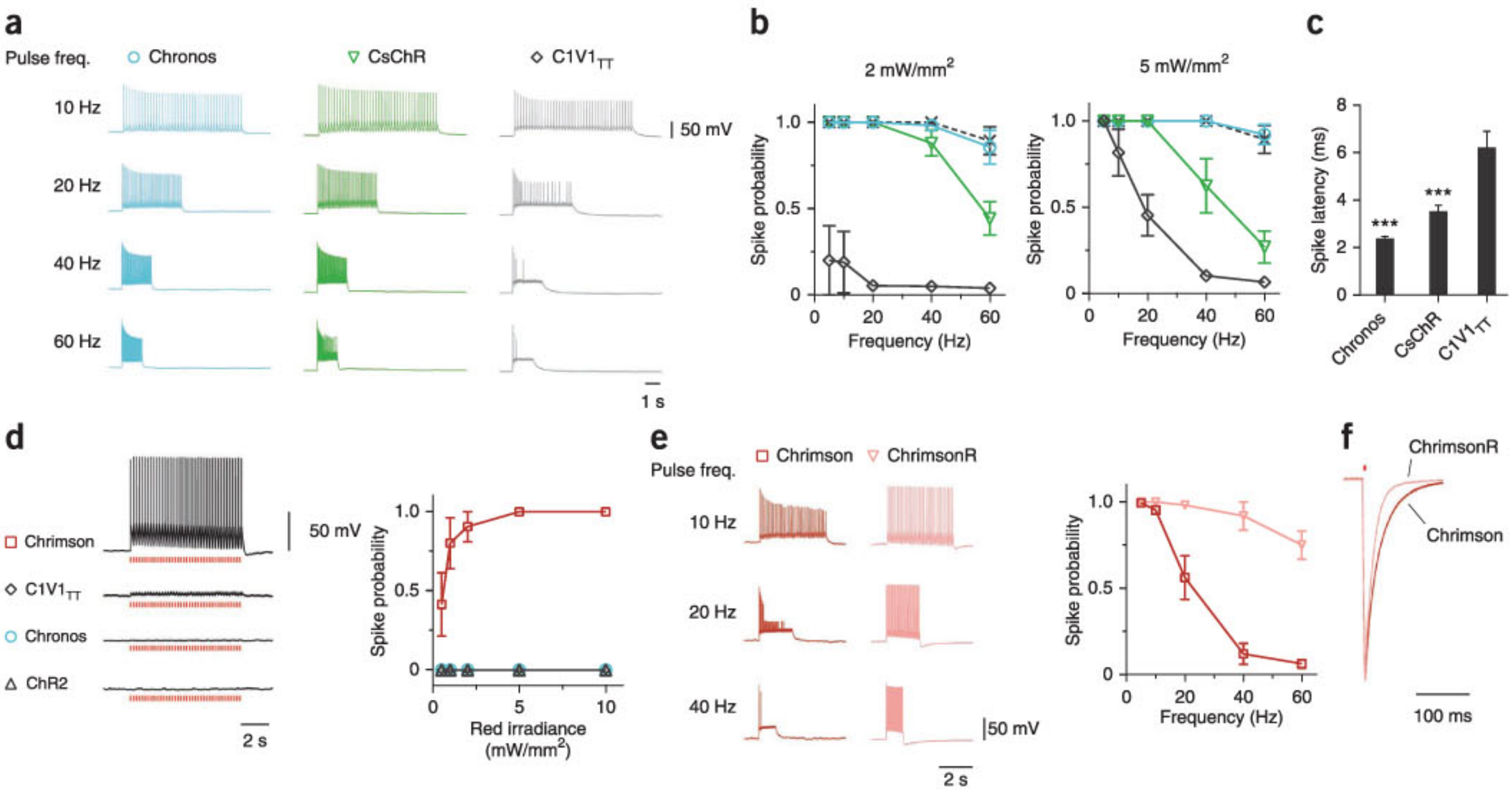


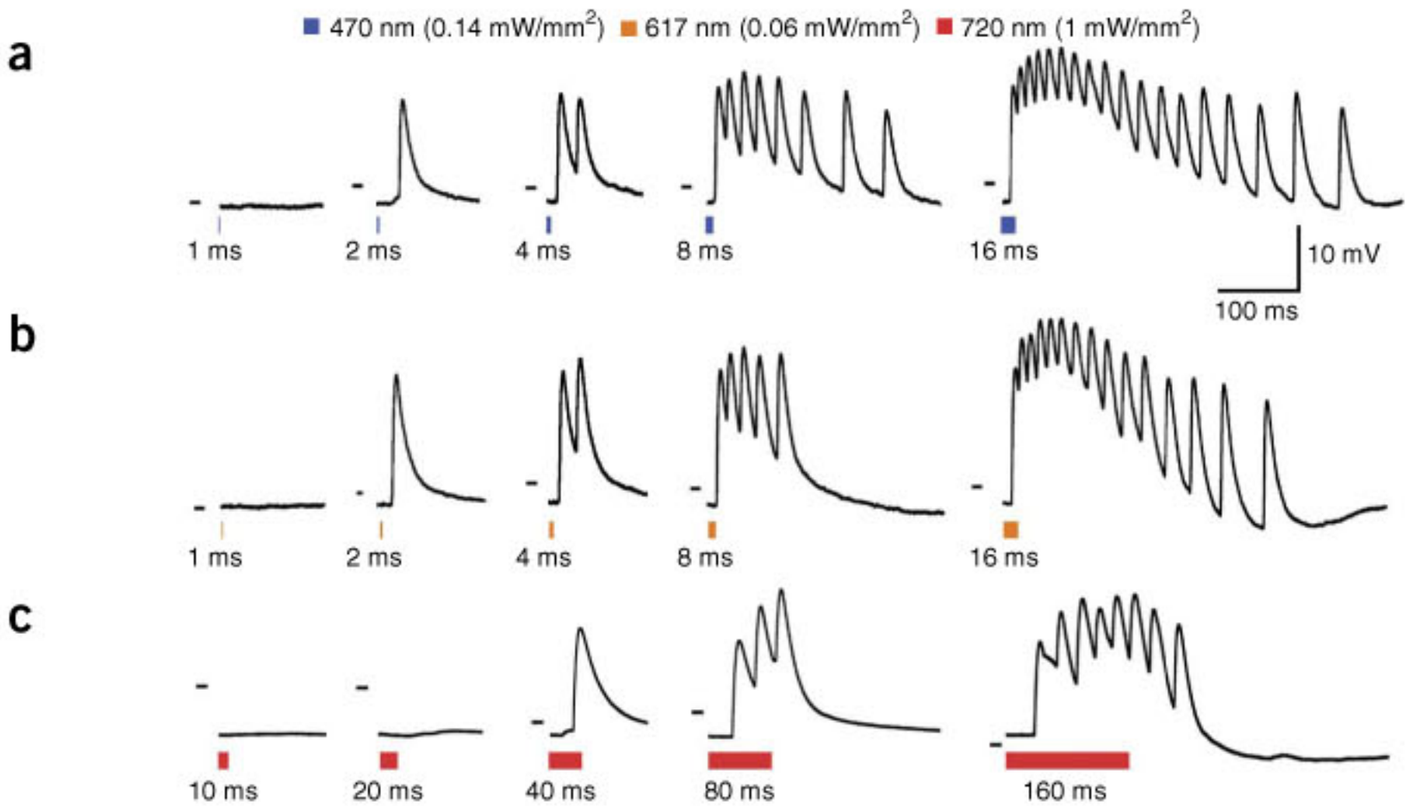
Figure 2 | Comparison of optical spiking in cultured neurons expressing different channelrhodopsins. (**a**–**c**) Green light–driven spiking fidelity. All green light spiking protocols used a train of 40 pulses, 2-ms pulse width, at 530 nm and at the indicated power (n = 5–8 cells for each opsin). (**a**) Representative green light–driven spiking traces at the indicated frequencies at 5 mW/mm^2. (**b**) Green light–driven spike probability over a range of frequencies. The dashed line is the electrical spiking control from Chronos-expressing neurons (this control comprised a train of 40 pulses at the indicated frequencies; each current-injection pulse was 5 ms long and was varied from 200 to 800 pA depending on each neuron's spike threshold). (**c**) Spike latencies (time between the light-pulse onset and the spike peak) calculated for 5-Hz trains at 5 mW/mm^2. (**d**–**f**) Comparison of spiking driven by red light (625 nm). (**d**) Representative current-clamp traces of red light response and spike fidelity (n = 5–8 cells for each opsin; 5-ms pulses, 5 Hz, 5 mW/mm^2). (**e**) Comparison of wild-type Chrimson and Chrimson K176R mutant (ChrimsonR) high-frequency red light spiking (n = 10 and 4 cells, respectively; 40-pulse train, 2-ms pulse width, 5 mW/mm^2). (**f**) Representative off-kinetics traces for Chrimson and ChrimsonR. ***P < 0.001; P = 0.0007 for CsChR and P < 0.0001 for Chronos; ANOVA with Dunnett's *post hoc* test with C1V1$_{TT}$ as reference. Plotted data are mean ± s.e.m.

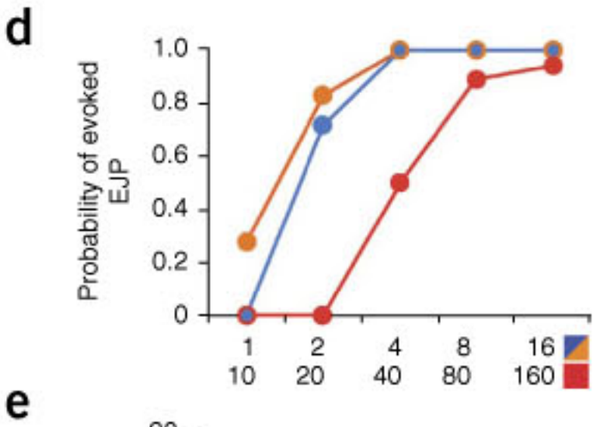
(720 nm) also triggered EJPs (**Fig. 3c**); however, high-intensity light (1 mW/mm^2) and long light-pulse durations (40–160 ms) were required for robust activation of the NMJ (**Fig. 3d**,**e**). As a control, we tested a commonly used first-generation ChR2-expressing transgenic fly (ref. 33) and examined responses to 470-nm and 617-nm light pulses. As in previous work, 470-nm light triggered EJPs in ChR2-expressing animals, but only after relatively long light pulses (16 ms); 617-nm light pulses did not trigger EJPs in ChR2 animals (**Supplementary Fig. 14**).

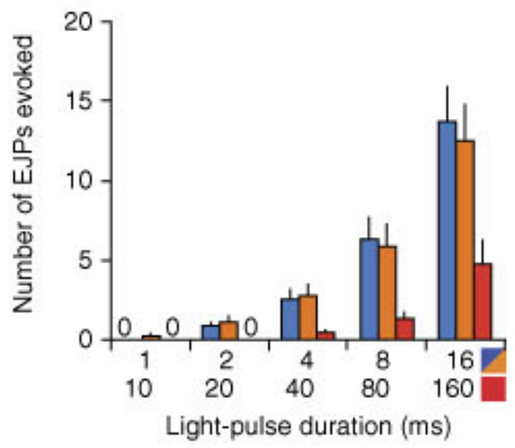
In adult flies, we expressed Chrimson in sweet taste receptor cells (using the Gr64f-Gal4 line to drive its expression in the proboscis and legs[34]) and measured the proboscis extension reflex (PER) elicited by light stimulation at different wavelengths and intensities (**Supplementary Video 1** and **Supplementary Fig. 15**). At 470 nm and 617 nm, PERs were robust at very low light intensities (0.02 mW/mm^2 and 0.015 mW/mm^2, respectively; **Fig. 3f**). Flies also responded to 720-nm light at an intensity substantially lower than that required in the larvae (0.07 mW/mm^2, 10 ms; **Fig. 3f** and **Supplementary Video 2**). 720 nm is believed to be outside of the fly photoreceptor light-absorption spectra[35,36]; we therefore hypothesized that Chrimson could be used without inducing visually driven behavioral artifacts. However, control flies showed a clear startle response to 720-nm stimuli in darkness (**Fig. 3h** and **Supplementary Video 3**). Nevertheless, we reasoned that the saliency of 720-nm light would drop if it were presented along with other visual stimuli at a wavelength well within the sensitivity of photoreceptors, as might be expected during visual behavior experiments in adult flies. As expected, the startle response was efficiently inhibited when we introduced flowing blue random dots during 720-nm stimulation (reduction from 93.2% nonzero startle responses, out of 44 valid trials in darkness, to 22.2% out of 45 in a circular arena displaying flowing dots; see Online Methods for statistics), and the PER of Gr64f-Gal4/Chrimson-expressing flies was preserved (**Fig. 3g** and **Supplementary Videos 4** and **5**). With a far red–shifted activation spectrum, Chrimson also allows direct brain stimulation without removing the cuticle while animals freely behave. We expressed Chrimson in a set of antennal lobe projection neurons (PNv-1 using the VT03194-Gal4 driver line) innervating the V glomerulus, which is known to respond to CO_2 and induces an avoidance response when activated[37]. In this nonvisual paradigm, flies in a circular arena reliably avoided quadrants lit by weak red light (617 nm, 0.015 mW/mm^2; **Supplementary Fig. 16** and **Supplementary Video 6**; paired t-test: P = 0.007; see Online Methods), whereas wild-type flies did not show a response (P = 0.502).

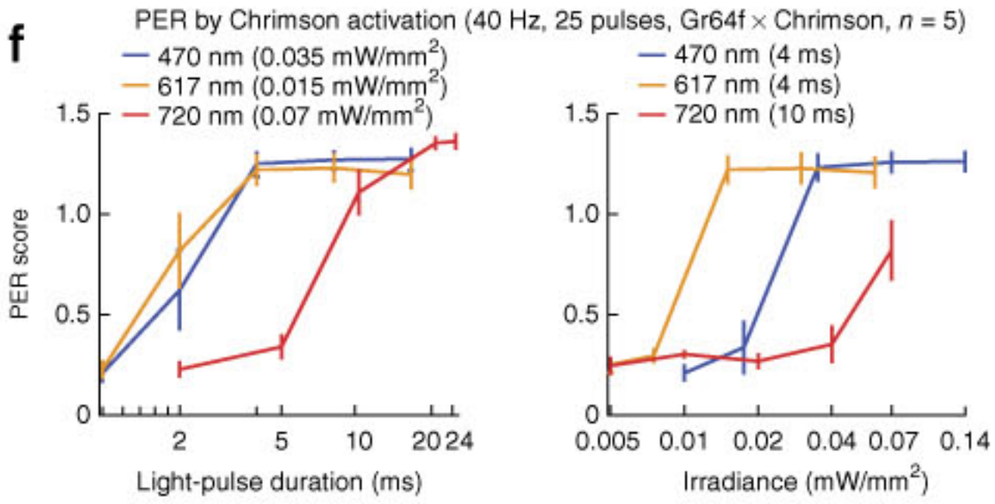
Principles of two-color independent neural excitation

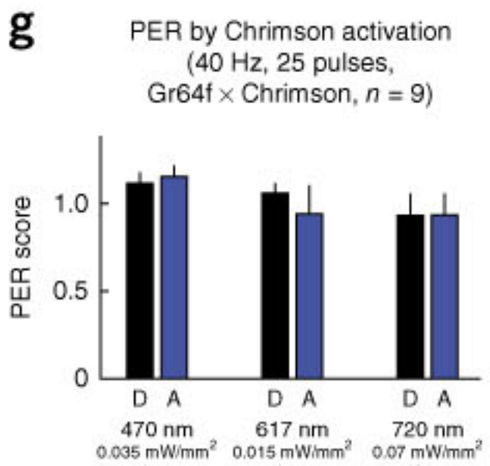
The fundamental limitation in creating an independent two-color channelrhodopsin pair is that all opsins can be driven to some extent by blue light. Additionally, neuron-to-neuron variation in opsin expression and optical scattering and absorption in tissue suggest that a large difference in effective blue light sensitivity











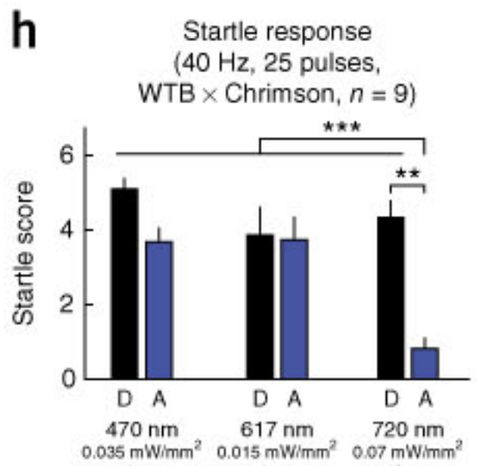


Figure 3 | Chrimson evokes action potentials in larval *Drosophila* motor neurons and triggers stereotyped behavior in adult *Drosophila*. (**a**–**c**) Intracellular recordings from m6 muscles in 3rd instar larvae expressing Chrimson in motor neurons. Responses to 470-nm, 617-nm and 720-nm light pulses of indicated power and increasing duration are shown. Dashes in each subpanel indicate −50 mV. (**d**) Probability of light-evoked excitatory junction potentials (EJPs) after pulses of 1, 2, 4, 8 or 16 ms in response to 470-nm and 617-nm light and after pulses of 10, 20, 40, 80 or 160 ms in response to 720-nm light. $n = 6$ muscles in 3 animals for all larvae experiments. (**e**) Mean number of EJPs evoked in response to light pulses. (**f**–**h**) Behavioral response of adult flies to light ($n = 5$ flies in each case). (**f**) Proboscis extension reflex (PER) of flies (pUAS-Chrimson-mVenus in attP18/w^-;*Gr64f-GAL4*/+; *Gr64f-GAL4*/+, shown as Gr64f × Chrimson) in response to 25 pulses of lights at 470 nm, 617 nm and 720 nm (see Online Methods for PER scoring). (**g**) PER of Gr64f × Chrimson flies to pulsed light in darkness (D) or in a visual arena with flowing blue random dots (A). (**h**) Startle response of control flies (pUAS-Chrimson-mVenus in attP18/+;+/+;+/+, shown as WTB × Chrimson) to visual stimuli as in **g**. The startle score is the number of moving legs after stimulation. ***$P < 0.001$, **$P < 0.01$. Error bars, s.e.m.

between blue and red-shifted channelrhodopsins, in addition to a large spectral separation, is required to guarantee robust, independent spiking in mammalian brain tissue. We here systematically explored channelrhodopsins' blue light sensitivity in cultured neurons, where, unlike in intact brain tissue, it is possible to precisely control light power.

In evaluating potential blue partners, we first examined the importance of fast channelrhodopsin kinetics. When Chrimson is exposed to blue light levels as dim as 0.1 mW/mm^2 over long durations, as a slow-to-activate blue channelrhodopsin might require, charge integration can result in action potentials (**Fig. 4a**). However, the on-kinetics of Chronos are roughly three times faster than those of ChR2 and ten times faster than those of Chrimson across all blue irradiances tested (**Fig. 4d**), which suggests it may be possible to activate Chronos without substantially activating Chrimson.

Thus, we examined expression variation in Chrimson cells (**Fig. 4b**), as this variance will translate into some cells exhibiting larger blue light depolarizations than others. We found cross-talk of up to 25 mV at typical blue light powers used for ChR2 excitation (1 mW/mm^2, 470 nm, 5-ms pulse). Given Chronos's high photocurrent (**Fig. 1c**), high light sensitivity (**Fig. 4c**) and fast on- and off-kinetics (**Figs. 1f,g** and **4d**), we examined whether these properties would translate into spiking at low blue irradiances. Chronos reliably drove 100% spiking at light powers as low as 0.05 mW/mm^2 and maintained this fidelity over two orders of magnitude to 20 mW/mm^2 (**Fig. 4e,f**). When we determined the minimum irradiance threshold to achieve 100% spiking (MIT_{100}), Chronos had a consistently lower MIT_{100} than ChR2 for similar GFP fluorescence levels (**Fig. 4f** and **Supplementary Fig. 17**), a result suggesting that Chronos's high effective light sensitivity is not due to higher expression and that Chronos can robustly mediate light-sensitive control of neural spiking across a range of expression levels (**Fig. 4f**) without altering neural excitability (**Fig. 4g**). These properties make Chronos an ideal blue candidate to be combined with Chrimson.

Validation of two-color excitation in brain slices

To determine acceptable blue and red irradiances to selectively drive Chronos and Chrimson without spiking cross-talk, we first examined the powers at which 470-nm and 625-nm LEDs could drive spiking in Chronos and Chrimson neurons as well as the

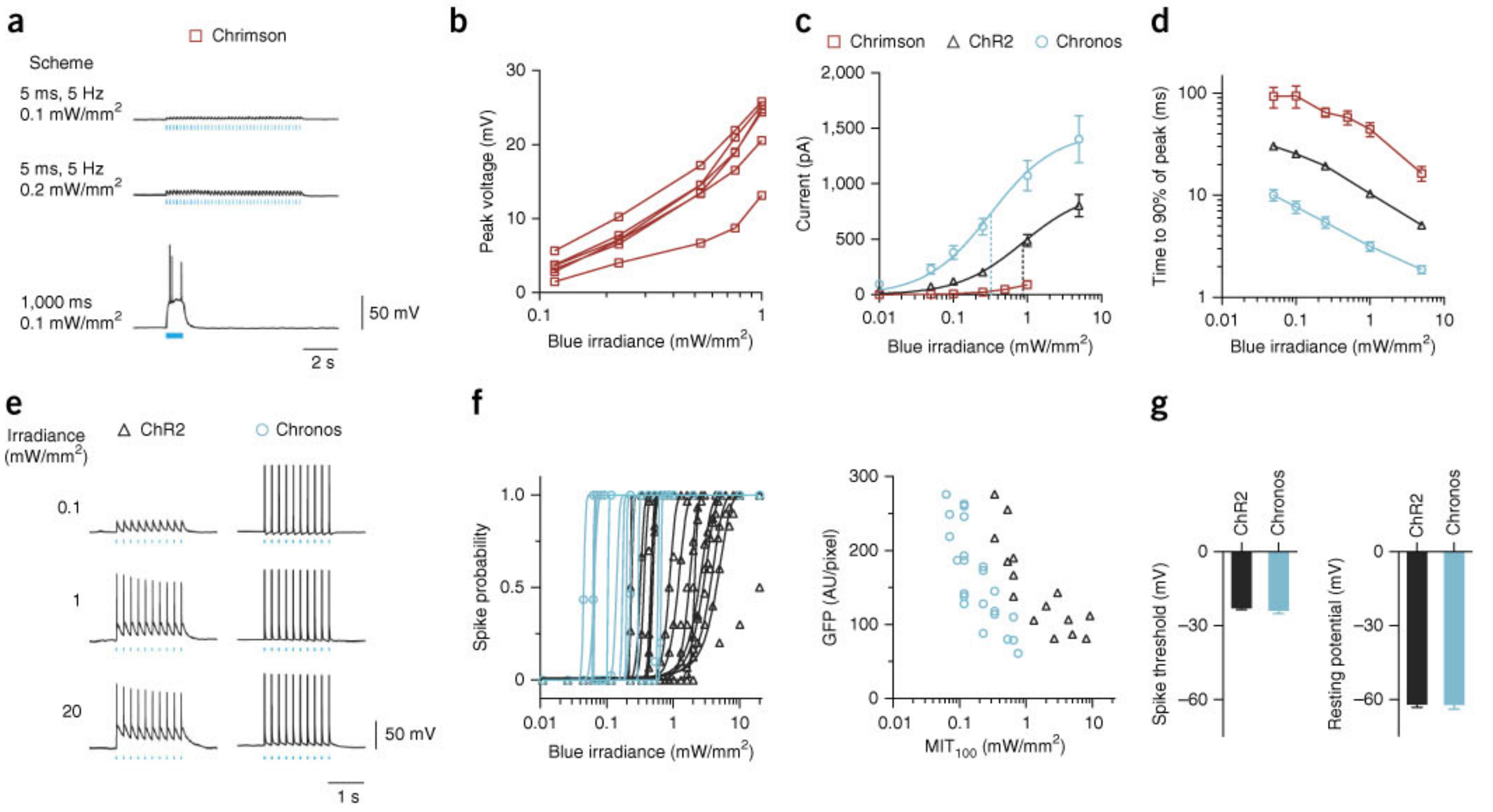


Figure 4 | Characterization of channelrhodopsin blue light (470-nm) sensitivities for two-color excitation in cultured neurons. (**a**) Current-clamp traces of representative Chrimson-expressing neuron under pulsed vs. continuous illumination. (**b**) Chrimson blue light–induced cross-talk voltages vs. irradiances for individual cells under pulsed illumination (5 ms, 5 Hz, $n = 5$ cells). (**c**) Photocurrent vs. blue irradiances (5-ms pulses; $n = 4$ cells for Chrimson, $n = 8$–10 cells for others). Vertical dashed lines indicate half-maximal points up the curves for ChR2 and Chronos as fitted. (**d**) Turn-on kinetics (1-s pulse; $n = 4$–7 cells; see **Supplementary Fig. 17b,c** for raw traces). (**e**–**g**) Comparison between ChR2 and Chronos spike probability over three logs of blue irradiance. All pulsed illuminations used 10 pulses, 5 Hz, 5 ms pulse width. (**e**) Representative spiking traces at the indicated irradiances. (**f**) Spike probability vs. blue light irradiance, plotted for individual Chronos- or ChR2-expressing neurons and minimum irradiance threshold for 100% spiking (MIT_{100}) as a function of GFP fluorescence. AU, arbitrary units. (**g**) Neuron spike threshold and resting potentials ($n = 16$–23 cells). Error bars, s.e.m.

conditions at which spike-level cross-talk occurred (**Fig. 5a–e**). We expressed Chrimson or Chronos in mouse cortical (layer 2/3) neurons via *in utero* electroporation and measured spike probabilities in opsin-expressing neurons in acute brain slices. As expected, we found that red light elicited spikes in only Chrimson-expressing cells (**Fig. 5c**) and blue light elicited spikes in Chronos-expressing cells using light powers as low as 0.05 mW/mm^2 (**Fig. 5d**). Chrimson-expressing neurons began to spike in response to blue light only when light powers were above 0.5 mW/mm^2 (**Fig. 5d**). This suggests that Chronos-expressing neurons can be driven with high fidelity without inducing spikes in Chrimson-expressing cells when using blue light irradiances between 0.2 and 0.5 mW/mm^2. Throughout this operational blue range, Chrimson-expressing cells showed subthreshold membrane depolarization events in response to blue light (**Fig. 5e**) comparable to the culture data (**Fig. 4b**).

We expressed Chrimson and Chronos in independent sets of neurons within the same cortical microcircuit in layer 2/3 neurons by electroporating Cre-on and Cre-off vectors[38]. To measure the synaptic output signal from the two targeted populations, we patch-clamped postsynaptic non-opsin-expressing neurons and measured optically evoked postsynaptic currents (PSCs) (**Fig. 5f–h**). For these experiments we chose a blue irradiance of 0.3 mW/mm^2 to elicit 100% Chronos spiking without Chrimson activation, and we used red irradiances of 1–4 mW/mm^2 to reliably activate Chrimson. We observed distinct PSC amplitudes when the cells were activated at the different wavelengths and light intensities (**Fig. 5i** and **Supplementary Figs. 18e** and **19a–c**). To examine whether these postsynaptic responses were cross-talk free at the synaptic transmission level, just as spiking events were at the action potential level, we singly expressed each opsin in distinct mouse brains, and we patch-clamped cells postsynaptic to opsin-expressing cells (**Fig. 5j–m** and **Supplementary Figs. 19d–g** and **20**). As expected, Chronos reliably drove synaptic events in response to blue light and never under red light (**Fig. 5k**; $P = 0.1$, paired t-test of peak current 30 ms before vs. 30 ms after red light delivery). Chrimson-expressing neurons reliably drove synaptic events upon red illumination and never upon blue illumination (**Fig. 5l**; $P = 0.43$, paired t-test of current before vs. after blue light delivery). Blue light–induced PSCs by Chrimson-expressing neurons began only outside this range, at 0.65 mW/mm^2 (**Fig. 5m**). We thus conclude it is possible to independently drive Chronos and Chrimson at the neural synaptic transmission level by using light powers determined by the operational blue irradiance range. As a final demonstration of Chronos's experimental utility, we found that pure axonal Chronos stimulation of retinal ganglion cell axons in the superior colliculus reliably elicited PSCs in downstream neurons (**Supplementary Fig. 21**).

DISCUSSION

We here present the results from a broad systematic screen of 61 algal opsins, sequenced *de novo* as part of a massive plant

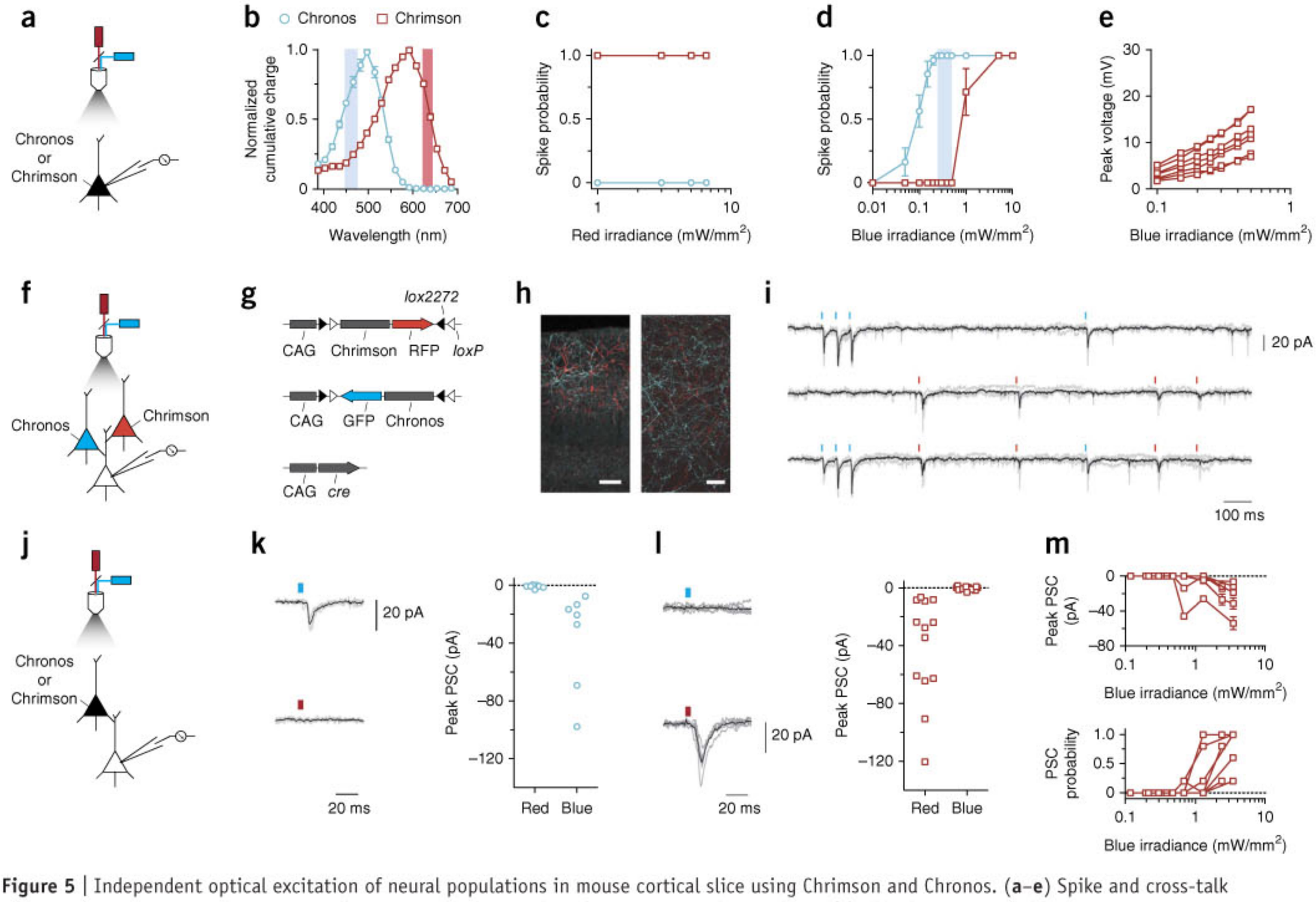


Figure 5 | Independent optical excitation of neural populations in mouse cortical slice using Chrimson and Chronos. (**a**–**e**) Spike and cross-talk characterization in opsin-expressing cells. Experimental optical configurations are depicted in **a**,**f**,**j**. (**b**) Chrimson and Chronos action spectra emphasizing (vertical shaded bars) the blue (470-nm) and red (625-nm) wavelengths used in this figure. (**c**–**e**) Current-clamp characterizations of Chrimson or Chronos expressing neurons in slice to determine optimal irradiance range for two-color excitation. Chrimson-GFP and Chronos-GFP were independently expressed in cortical layer 2/3 neurons in separate mice. 5-ms, 5-Hz light pulses were used; $n = 7$ cells from 3 animals for Chrimson; $n = 11$ cells from 4 animals for Chronos. (**c**,**d**) Spike probability vs. irradiance for red (**c**) and blue light (**d**). The blue vertical shaded bar represents the blue irradiance range where Chronos drove spikes at 100% probability and no cross-talk spike was ever observed for any Chrimson neurons. (**e**) Chrimson subthreshold cross-talk voltage in individual neurons vs. blue irradiances; compare to **Figure 4b**. (**f**–**i**) Postsynaptic currents (PSCs) in non-opsin-expressing neurons downstream of Chrimson and Chronos expressing neurons in brain slice with both opsins introduced into separate neural populations. Stimulation parameters: 0.3 mW/mm^2 for blue, 4 mW/mm^2 for red, 5-ms pulses; 6 neurons from 3 animals. All synaptic transmission slice experiments were done using wide-field illumination (**Supplementary Fig. 18**). (**g**) Triple-plasmid electroporation scheme for mutually exclusive Chrimson and Chronos expression in different sets of layer 2/3 cortical pyramidal cells. (**h**) Histology of intermingled Chrimson- (red) and Chronos-expressing (blue) neurons in layer 2/3 (left, taken at 10× magnification) and their axons (right, taken at 60× magnification). Scale bars: 100 µm (left), 20 µm (right). (**i**) PSCs in response to optical Poisson stimulation with blue and red light; shown are raw voltage traces (gray) with average trace (black) from a single neuron experiencing blue (top), red (center) or both (bottom) light pulses. (PSC traces from neurons downstream of mutually exclusive Chrimson- and Chronos-expressing neurons in response to blue or red light are in **Supplementary Fig. 18e**.) (**j**–**m**) PSCs in non-opsin-expressing neurons downstream of Chronos- or Chrimson-expressing neurons. Conditions are as in **f**–**i**, except pulses were delivered at 0.2 Hz. $n = 7$ cells from 2 animals for Chronos; $n = 12$ cells from 4 animals for Chrimson. The black trace is the averaged response; gray traces are individual trials. (**k**) Chronos-driven PSCs under blue or red light, obtained from a representative neuron (left), with population data (right). (**l**) Chrimson-driven PSCs under blue or red light traces, obtained from a representative neuron (left), with population data (right). (**m**) Chrimson-driven PSC amplitudes (top) and the probability of observing a PSC at all (bottom) vs. blue irradiances.

transcriptome endeavor[20] (http://www.onekp.com/). We discovered Chronos, an ultra-light-sensitive blue channelrhodopsin with faster kinetics than previously described opsins and which might represent an excellent general-use channelrhodopsin. We also discovered Chrimson, a red light–drivable channelrhodopsin 45 nm more red shifted than any previous channelrhodopsin and which might be useful in scenarios in which red light stimulation is essential. Previous channelrhodopsins that can be driven off-peak by red light (630 nm) have on-off kinetics exceeding 100 ms (refs. 9,14): perhaps a limitation of their original VChR1 scaffold. The spectral and kinetic properties of Chrimson and Chronos will enable fundamentally new types of experiments.

We explored the utility of Chrimson with red light in *D. melanogaster*. Chrimson was able to mediate responses in larval and adult flies with extremely low light powers across all wavelengths tested, most likely owing to the high expression achieved with the expression cassette used (see Online Methods). Further, light-induced startle responses of adult flies were significantly reduced at longer wavelengths (720 nm) when visual distractors of shorter wavelength were provided. Thus, Chrimson may be

useful for temporally precise neuronal stimulation in *Drosophila* behavioral experiments. Curiously, our results suggest that the photosensitivity of the fly eye extends well beyond the wavelength of typical room light for fly experiments (~650 nm), highlighting the utility of the 720-nm stimulation protocol we introduce for a wider range of behavioral experiments. Wavelengths lower than 720 nm will be useful in situations in which startle responses do not affect measurements of the parameters under study, such as stimulating neurons in the brain through intact cuticle to induce nonvisually driven behaviors as here shown with 617-nm light and CO_2-responding neurons, something that has not yet been reliably demonstrated using optogenetic techniques. The ability to perform optogenetic behavioral experiments in intact flies with Chrimson at redder wavelengths than ReaChR[14,29], due to the ~45-nm spectral shift, may permit lower light levels to be used and reduce visual system–mediated artifacts.

Using both Chronos and Chrimson, we found an ample blue light irradiance range that evoked reliable Chronos-induced spikes without any Chrimson-induced spikes in mouse cortical slice and that also allowed Chronos-induced synaptic transmission without any Chrimson-induced synaptic transmission. As most two-color experimental setups will not be identical to our slice demonstration, several constraints exist upon the use of effective blue light sensitivity to achieve two-color separation. Whenever possible, Chronos should be expressed under the stronger promoter and the more excitable neuron type to minimize Chrimson-induced depolarization. The experiments in this paper characterized two-color excitation with symmetric promoters and cell types to stringently test whether Chronos and Chrimson can, without exploiting differences in cellular excitability, present a clear separation due to biophysical properties. Neuroscientists seeking to manipulate two different cell populations may find that expressing Chronos at high levels in the more excitable cell type may further increase the reliable blue dynamic range.

Although Chronos has the fastest kinetic properties among reported channelrhodopsins, blue light pulses delivered at very high frequencies as part of a two-color experiment could in principle lead to charge integration and, thus, to Chrimson blue spiking cross-talk. The temporal precision of Chronos- versus Chrimson-mediated synaptic events may also depend on light power, potentially limiting usage in scenarios requiring submillisecond timing of synaptic release, although ~1-ms jitter is achievable. We have additionally observed in postsynaptic experiments that the stimulation frequency for both red and blue pulses is fundamentally limited by the wild-type Chrimson, the slowest of the opsin pair. This is why we engineered ChrimsonR, a faster Chrimson kinetic mutant, which has similar blue light sensitivity to the wild type but allows modulation at above the 20-Hz range. It may be desirable to further improve Chrimson's kinetics and decrease its light sensitivity to enable high-frequency modulation in both the red and blue channels and to increase the usable blue irradiance range.

For applications *in vivo*, it would be important to illuminate the circuit region of interest with powers that fall within the windows here defined; for mammals, one might use alternative illumination methods such as three-dimensional optical waveguides[39] or wireless LED implants[40]. Although these constraints add complexity to experiments, they may also enable *in vivo* two-color experiments not previously possible.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. GenBank/EMBL/DDBJ: new channelrhodopsins identified here are listed under accession codes KF992030–KF992090; see also **Supplementary Table 1**.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank A. Karpova (Janelia Farm) for technical advice, reagents and generous assistance with construct preparation for *Drosophila*; K. Hibbard and members of the Janelia Fly Core for fly husbandry and assistance with fly crosses; and J. Pulver for technical advice and assistance with data analysis software. We thank Y. Aso, W. Ming and G. Rubin (Janelia Farm) for kindly allowing us to use their circular light arena and for useful discussion. We thank I. Negrashov, S. Sawtelle and J. Liu for arena-related development and support. We thank J.R. Carlson (Yale University) for Gr64f-Gal4 flies, W.D. Tracey Jr. (Duke University) for UAS-Chr2 flies, G.M. Rubin (Janelia Farm) for pBDP-Gal4 flies and B.J. Dickson (IMP, Vienna and Janelia Farm) for VT031497-Gal4 flies.

S.S.K., S.R.P. and V.J. were supported by the Howard Hughes Medical Institute. The 1000 Plants (1KP) initiative, led by G.K.-S.W., is funded by the Alberta Ministry of Enterprise and Advanced Education, Alberta Innovates Technology Futures (AITF) Innovates Centre of Research Excellence (iCORE), Musea Ventures, and BGI-Shenzhen. B.Y.C. and E.S.B. were funded by Defense Advanced Research Projects Agency (DARPA) Living Foundries HR0011-12-C-0068. B.Y.C. was funded by the US National Science Foundation (NSF) Biophotonics Program. M.C.-P. was funded by US National Institutes of Health (NIH) grant 5R01EY014074-18. E.S.B. was funded by the MIT Media Lab, Office of the Assistant Secretary of Defense for Research and Engineering, Harvard/MIT Joint grants in Basic Neuroscience, NSF (especially CBET 1053233 and EFRI 0835878), NIH (especially 1DP20D002002, 1R01NS067199, 1R01DA029639, 1R01GM104948, 1RC1MH088182 and 1R01NS075421), Wallace H. Coulter Foundation, Alfred P. Sloan Foundation, Human Frontiers Science Program, New York Stem Cell Foundation Robertson Neuroscience Investigator Award, Institution of Engineering and Technology A.F. Harvey Prize, and Skolkovo Institute of Science and Technology.

AUTHOR CONTRIBUTIONS

N.C.K., E.S.B., M.C.-P. and V.J. contributed to the study design and data analysis. G.K.-S.W. and B.Y.C. oversaw transcriptomic sequencing. E.S.B. and M.C.-P. supervised mammalian opto/electrophysiological parts of the project. N.C.K. coordinated all experiments and data analysis. N.C.K., Y.K.C., A.S.C. and T.K.M. conducted and analyzed all *in vitro* electrophysiology. M.M., B.S., N.C.K., T.K.M., E.J.C., Z.T., J.W., Y.X., Z.Y. and Y.Z. conducted algal RNA experiments or transcriptome sequencing and analysis. N.C.K., Y.M. and A.B.-B. performed and analyzed all slice electrophysiology. V.J. prepared Chrimson for injection into *Drosophila*. S.S.K. and V.J. designed adult fly behavior experiments. S.S.K. performed all fly behavior experiments and data analysis. S.R.P. designed, performed and analyzed all larval *Drosophila* experiments. Correspondence should be addressed to V.J. (vivek@janelia.hhmi.org) for Chrimson flies. All authors contributed to the discussions and writing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Boyden, E.S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005).
2. Han, X. & Boyden, E.S. Multiple-color optical activation, silencing, and desynchronization of neural activity, with single-spike temporal resolution. *PLoS ONE* **2**, e299 (2007).
3. Chow, B.Y. *et al.* High-performance genetically targetable optical neural silencing by light-driven proton pumps. *Nature* **463**, 98–102 (2010).
4. Zhang, F. *et al.* Multimodal fast optical interrogation of neural circuitry. *Nature* **446**, 633–639 (2007).

5. Gradinaru, V. *et al.* Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* **141**, 154–165 (2010).
6. Boyden, E.S. A history of optogenetics: the development of tools for controlling brain circuits with light. *F1000 Biol. Rep.* **3**, 11 (2011).
7. Zhang, F. *et al.* Red-shifted optogenetic excitation: a tool for fast neural control derived from Volvox carteri. *Nat. Neurosci.* **11**, 631–633 (2008).
8. Erbguth, K., Prigge, M., Schneider, F., Hegemann, P. & Gottschalk, A. Bimodal activation of different neuron classes with the spectrally red-shifted channelrhodopsin chimera C1V1 in *Caenorhabditis elegans*. *PLoS ONE* **7**, e46827 (2012).
9. Yizhar, O. *et al.* Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477**, 171–178 (2011).
10. Prigge, M. *et al.* Color-tuned channelrhodopsins for multiwavelength optogenetics. *J. Biol. Chem.* **287**, 31804–31812 (2012).
11. Wang, W. *et al.* Tuning the electronic absorption of protein-embedded all-trans-retinal. *Science* **338**, 1340–1343 (2012).
12. Waddell, W.H., Schaffer, A.M. & Becker, R.S. Visual pigments. 3. Determination and interpretation of the fluorescence quantum yields of retinals, Schiff bases, and protonated Schiff bases. *J. Am. Chem. Soc.* **95**, 8223–8227 (1973).
13. Govorunova, E.G., Spudich, E.N., Lane, C.E., Sineshchekov, O.A. & Spudich, J.L. New channelrhodopsin with a red-shifted spectrum and rapid kinetics from *Mesostigma viride*. *mBio* **2**, e00115–e00111 (2011).
14. Lin, J.Y., Knutsen, P.M., Muller, A., Kleinfeld, D. & Tsien, R.Y. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat. Neurosci.* **16**, 1499–1508 (2013).
15. Kleinlogel, S. *et al.* Ultra light-sensitive and fast neuronal activation with the Ca^{2+}-permeable channelrhodopsin CatCh. *Nat. Neurosci.* **14**, 513–518 (2011).
16. Berndt, A., Yizhar, O., Gunaydin, L.A., Hegemann, P. & Deisseroth, K. Bi-stable neural state switches. *Nat. Neurosci.* **12**, 229–234 (2009).
17. Bamann, C., Gueta, R., Kleinlogel, S., Nagel, G. & Bamberg, E. Structural guidance of the photocycle of channelrhodopsin-2 by an interhelical hydrogen bond. *Biochemistry* **49**, 267–278 (2010).
18. Mattis, J. *et al.* Principles for applying optogenetic tools derived from direct comparative analysis of microbial opsins. *Nat. Methods* **9**, 159–172 (2012).
19. Govorunova, E.G., Sineshchekov, O.A., Li, H., Janz, R. & Spudich, J.L. Characterization of a highly efficient blue-shifted channelrhodopsin from the marine alga *Platymonas subcordiformis*. *J. Biol. Chem.* **288**, 29911–29922 (2013).
20. Johnson, M.T. *et al.* Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS ONE* **7**, e50226 (2012).
21. Lin, J.Y. A user's guide to channelrhodopsin variants: features, limitations and future developments. *Exp. Physiol.* **96**, 19–25 (2011).
22. Nagel, G. *et al.* Light activation of channelrhodopsin-2 in excitable cells of *Caenorhabditis elegans* triggers rapid behavioral responses. *Curr. Biol.* **15**, 2279–2284 (2005).
23. de Vries, S.E. & Clandinin, T.R. Loom-sensitive neurons link computation to action in the *Drosophila* visual system. *Curr. Biol.* **22**, 353–362 (2012).
24. Schroll, C. *et al.* Light-induced activation of distinct modulatory neurons triggers appetitive or aversive learning in *Drosophila* larvae. *Curr. Biol.* **16**, 1741–1747 (2006).
25. Gaudry, Q., Hong, E.J., Kain, J., de Bivort, B.L. & Wilson, R.I. Asymmetric neurotransmitter release enables rapid odour lateralization in *Drosophila*. *Nature* **493**, 424–428 (2013).
26. Honjo, K., Hwang, R.Y. & Tracey, W.D. Jr. Optogenetic manipulation of neural circuits and behavior in *Drosophila* larvae. *Nat. Protoc.* **7**, 1470–1478 (2012).
27. Zhang, W., Ge, W. & Wang, Z. A toolbox for light control of *Drosophila* behaviors through Channelrhodopsin 2-mediated photoactivation of targeted neurons. *Eur. J. Neurosci.* **26**, 2405–2416 (2007).
28. Xiang, Y. *et al.* Light-avoidance-mediating photoreceptors tile the *Drosophila* larval body wall. *Nature* **468**, 921–926 (2010).
29. Inagaki, H.K. *et al.* Optogenetic control of *Drosophila* using a red-shifted channelrhodopsin reveals experience-dependent influences on courtship. *Nat. Methods* doi:10.1038/nmeth.2765 (22 December 2013).
30. Claridge-Chang, A. *et al.* Writing memories with light-addressable reinforcement circuitry. *Cell* **139**, 405–415 (2009).
31. Pulver, S.R., Pashkovski, S.L., Hornstein, N.J., Garrity, P.A. & Griffith, L.C. Temporal dynamics of neuronal activation by Channelrhodopsin-2 and TRPA1 determine behavioral output in *Drosophila* larvae. *J. Neurophysiol.* **101**, 3075–3088 (2009).
32. Bernstein, J.G., Garrity, P.A. & Boyden, E.S. Optogenetics and thermogenetics: technologies for controlling the activity of targeted cells within intact neural circuits. *Curr. Opin. Neurobiol.* **22**, 61–71 (2012).
33. Hwang, R.Y. *et al.* Nociceptive neurons protect *Drosophila* larvae from parasitoid wasps. *Curr. Biol.* **17**, 2105–2116 (2007).
34. Dahanukar, A., Lei, Y.T., Kwon, J.Y. & Carlson, J.R. Two *Gr* genes underlie sugar reception in *Drosophila*. *Neuron* **56**, 503–516 (2007).
35. Minke, B. & Kirschfeld, K. The contribution of a sensitizing pigment to the photosensitivity spectra of fly rhodopsin and metarhodopsin. *J. Gen. Physiol.* **73**, 517–540 (1979).
36. Salcedo, E. *et al.* Blue- and green-absorbing visual pigments of *Drosophila*: ectopic expression and physiological characterization of the R8 photoreceptor cell-specific Rh5 and Rh6 rhodopsins. *J. Neurosci.* **19**, 10716–10726 (1999).
37. Lin, H.H., Chu, L.A., Fu, T.F., Dickson, B.J. & Chiang, A.S. Parallel neural pathways mediate CO_2 avoidance responses in *Drosophila*. *Science* **340**, 1338–1341 (2013).
38. Atasoy, D., Aponte, Y., Su, H.H. & Sternson, S.M. A FLEX switch targets Channelrhodopsin-2 to multiple cell types for imaging and long-range circuit mapping. *J. Neurosci.* **28**, 7025–7030 (2008).
39. Zorzos, A.N., Scholvin, J., Boyden, E.S. & Fonstad, C.G. Three-dimensional multiwaveguide probe array for light delivery to distributed brain circuits. *Opt. Lett.* **37**, 4841–4843 (2012).
40. Kim, T.I. *et al.* Injectable, cellular-scale optoelectronics with applications for wireless optogenetics. *Science* **340**, 211–216 (2013).

Addendum: Digestion and depletion of abundant proteins improves proteomic coverage

Bryan R Fonslow, Benjamin D Stein, Kristofor J Webb, Tao Xu, Jeong Choi, Sung Kyu Park & John R Yates III

Nature Methods 10, 54–56 (2013); published online 18 November 2013; addendum published after print 27 February 2014

Recently we reported improved proteome coverage and quantitation metrics for low-abundance proteins within whole proteomes by implementing a digestion and depletion strategy (DigDeAPr) before a standard shotgun proteomic analysis. Our goal was to improve the detection of low-abundance proteins by reducing the proteolytic background of highly sampled peptides derived from high-abundance proteins. We rationalized that our gains in proteome coverage resulted from the selective digestion and removal of abundant proteins as peptides. Since the publication of the method, the mechanism by which our gains were achieved was challenged in a Correspondence by Ye *et al.*[1]. In response, we have reanalyzed our data in a peptide-centric manner and propose a refined kinetic mechanism consistent with established competitive substrate kinetics.

Through a simplified derivation beginning with a classical Michaelis-Menten competitive-substrate model and further quantitative analysis of our data, we provide a refined depletion mechanism that more accurately describes the complex mixtures we previously analyzed. Our revised qualitative expression describing depletion of early generated peptides from proximal fast tryptic cleavage sites with high specificity constants (V/K) (**Supplementary Note 1**) is illustrated by the following equation

$$\chi_{\mathrm{A,depleted}} = \frac{(\chi_{\mathrm{A}})_{t=0}\, e^{-\left(\frac{V}{K}\right)_{\mathrm{A}} t_{\mathrm{c}}}}{\sum (\chi_n)_{t=0}\, e^{-\left(\frac{V}{K}\right)_n t_{\mathrm{c}}}} - \frac{(\chi_{\mathrm{A}})_{t=0}\, e^{-\left(\frac{V}{K}\right)_{\mathrm{A}} t_{\mathrm{d}}}}{\sum (\chi_n)_{t=0}\, e^{-\left(\frac{V}{K}\right)_n t_{\mathrm{d}}}} \quad (1)$$

where $\chi_{\mathrm{A,depleted}}$ is the mole fraction of substrate A after complete (t_{c}) and depletion (t_{d}) digestion times expressed as mole fractions of total substrate cleavage sites. So expressed, tryptic sites have different specificity constants as well as abundances. Substrate cleavage results in the generation of two shorter polypeptides that can be subsequently cleaved into more substrates over time. The relative cleavage rates are governed by each site's relative specificity constant. From this perspective, we redefine the mechanism for depletion and enrichment of the DigDeAPr method. Early generated peptides, derived from fast substrate sites (i.e., those with high V/K) within ~100 amino acids of each other, are removed during our 10-kDa molecular-weight-cutoff spin-filter depletion step. The clearing of these early generated peptides before further digestion allows enrichment of peptides resulting from slower tryptic sites in the subsequent complete digestion step.

Using equation (1) we illustrate the expected adjustment in peptide abundance resulting from limited digestion and depletion (**Fig. 1a**) as driven by the relative cleavage-site specificity constants (V/K). When peptide abundance is considered between control and DigDeAPr runs, the expected trend is observed (**Fig. 1b** and **Supplementary Fig. 1**), a result consistent with our revised digestion and depletion theory. Notably, the use of tenfold more starting material and depletion of early generated peptides equalized the measured abundance of all peptides (**Supplementary Note 2** and

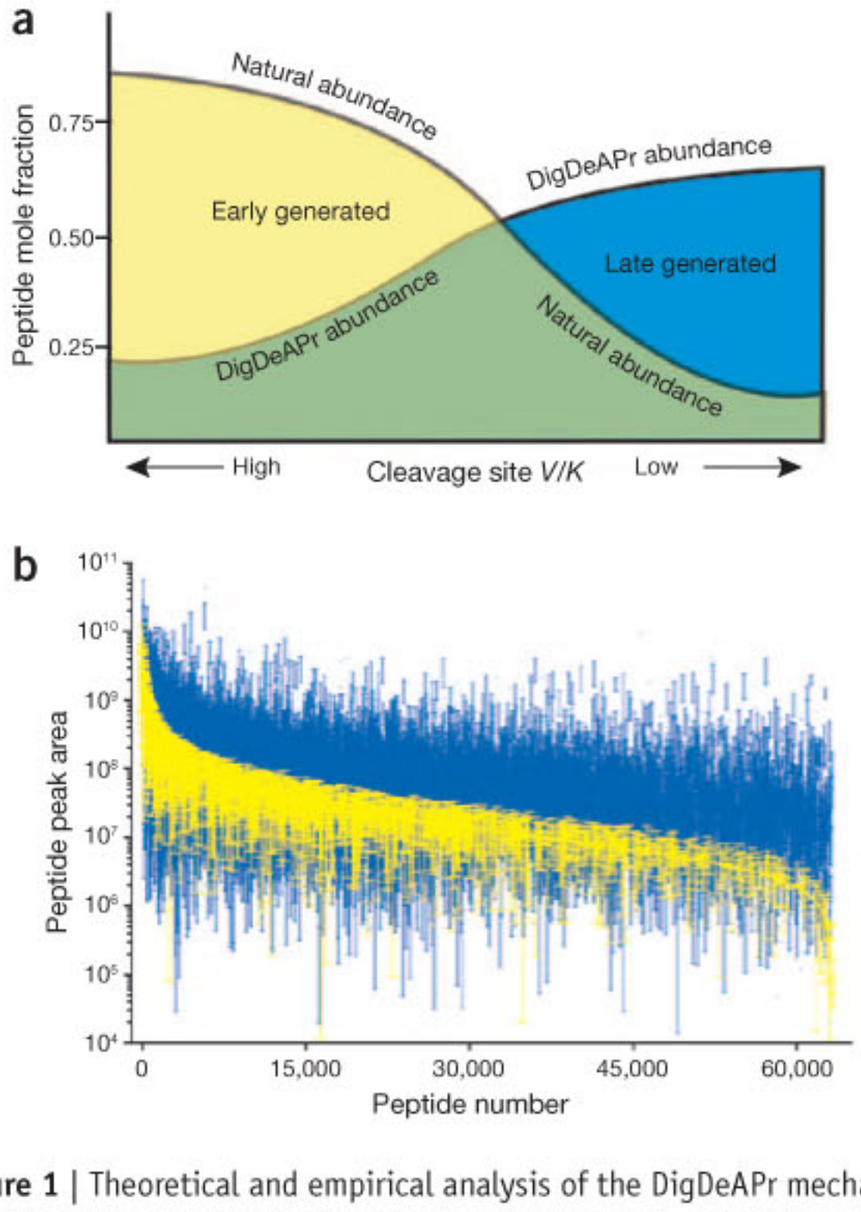


Figure 1 | Theoretical and empirical analysis of the DigDeAPr mechanism. (**a**) Schematic of our refined mechanism for digestion and depletion based on the cleavage-site specificity constant (V/K) for a given protease. The natural abundances of peptides from a complete protease digestion are adjusted by the use of ten times as much material and depletion of early generated peptides to enrich for late generated peptides with lower cleavage-site specificity constants. (**b**) Rank-abundance plot of peptide chromatographic peak areas from triplicate control (yellow) and DigDeAPr (blue) runs representing early and late generated peptides, respectively. Error bars, s.d.

Supplementary Figs. 2–4). Because peptide abundances are used to estimate protein abundance with shotgun proteomics[2–5], the equalization of peptides also equalizes the measurable abundance of proteins, as we found empirically in our initial analysis.

Our DigDeAPr runs provide a defined, limited digestion time point for consideration of the aforementioned kinetic efficiencies through analysis of early and late generated peptides and fast and slow tryptic cleavage sites (**Supplementary Note 3**). Early generated peptides should be depleted and have lower abundances after DigDeAPr when compared to control runs, whereas late generated peptides should be enriched and have higher abundances. Using label-free chromatographic peak-area ratios of peptides in both control and DigDeAPr runs, we quantified 13,628 and 13,112 peptides in human embryonic kidney (HEK) cells (**Fig. 2a**) and yeast cells, respectively, that were used to classify peptides as early or late generated by their relative ratios. Both distributions showed defined populations of peptides that were depleted ($\log_2$ ratio ≤ -1), unchanged ($-1 < \log_2$ ratio < 1) and enriched ($\log_2$ ratio ≥ 1). Focusing on the HEK peptide distribution, motif analysis of cleaved

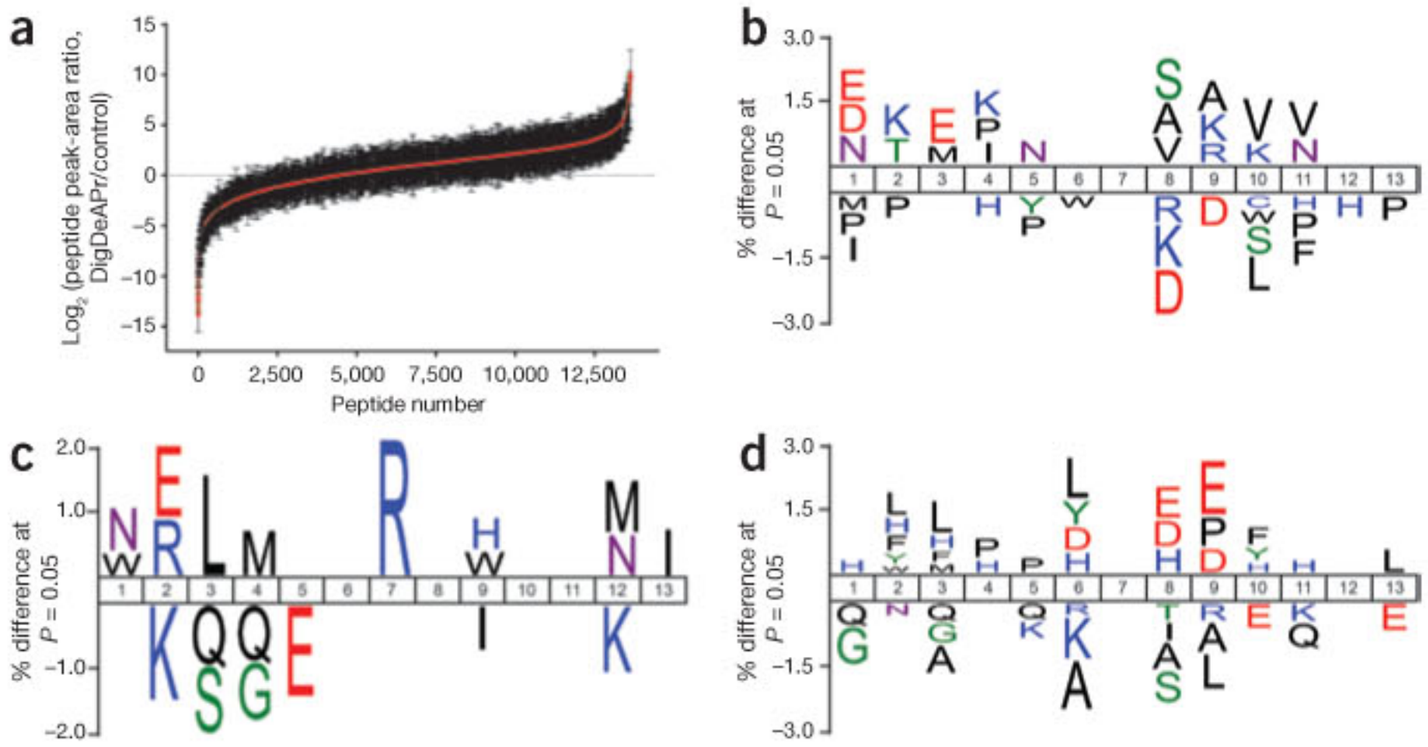


Figure 2 | Motif analysis to support the DigDeAPr mechanism. (**a**) Distribution of quantified HEK peptide ratios using label-free peak-area measurements (Online Methods). (**b**–**d**) Tryptic-site motif analysis using iceLogo[7] and peptide ratios from the analysis of HEK cells, categorized by peak-area ratios: depleted ($\log_2$ ratio ≤ −1) cleaved sites (n = 5,834) versus unchanged (−1 < $\log_2$ ratio < 1) cleaved sites (n = 11,885) (**b**), depleted missed cleavage sites (n = 2,846) versus unchanged missed cleavage sites (n = 5,438) (**c**), and enriched ($\log_2$ ≥ 1) cleaved sites (n = 22,239) versus unchanged cleaved sites (n = 11,885) (**d**).

(**Fig. 2b**) and missed cleaved (**Fig. 2c**) tryptic sites on depleted peptides confirmed that early generated peptides from proximal fast tryptic cleavage sites (<~100 amino acids apart) were selectively removed during the 10-kDa depletion step (**Supplementary Note 4**). Similarly, tryptic motifs of enriched, late generated peptides represent slow cleavage sites (**Fig. 2d**) that remained uncleaved within polypeptides of >10 kDa at the depletion time point. Thus, consideration of tryptic sites and peptides in the digestion and depletion mechanism is essential and illustrates the depletion and enrichment of peptides from fast and slow tryptic cleavage sites, respectively.

By considering these early and late generated peptides in our protein abundance analyses, we notably still observed an abundance-based depletion and enrichment trend in both yeast and HEK cells: higher-abundance proteins have more early generated peptides identified, and lower- abundance proteins have more late generated peptides identified (**Supplementary Fig. 5** and **Supplementary Note 5**). On the basis of these data and our understanding of peptide sampling in shotgun proteomics[2], we conclude that our gains originate from analysis of a different population of enriched, late generated peptides. That is, depletion of early generated peptides from high-abundance proteins removes enough proteolytic background to unmask and identify more late generated peptides from low-abundance proteins. Although we may not have explicitly depleted abundant proteins through digestion, in our reanalysis we found that depletion or enrichment of single peptides accounted for ~30% (1/slope = 0.298) of the observed protein abundance depletion or enrichment, respectively, explained by ~60% (coefficient of determination R^2 = 0.57) of the protein abundance measurements (**Supplementary Fig. 6** and **Supplementary Note 6**). Additionally, we found a notable overlap in depleted, early generated yeast peptides and 'proteotypic' yeast peptides (**Supplementary Fig. 7** and **Supplementary Note 5**). Although proteotypic peptides can be used to robustly identify and quantify many proteins, they can also act as proteolytic background for other less abundant or less sampled proteins and peptides[6]. Our results collectively indicate that depletion of highly sampled, abundant, easily identified, proteotypic peptides has a similar effect as depleting abundant proteins to improve identification and quantification of peptides from low-abundance proteins.

With our reexamined view of peptide abundance changes and their correlation to protein changes, we propose a refined mechanism by which our proteome coverage and quantitation gains are realized through digestion and depletion: depletion of early generated peptides and enrichment of late generated peptides equalizes measurable peptide abundances and unmasks less proteotypic peptides for improvements in low-abundance protein identification and quantification. We suggest that DigDeAPr should represent digestion and depletion of abundantly sampled peptides and proteins through enrichment of less easily digested and identifiable proteins and peptides. Nonetheless, the combination of tenfold more starting material with limited digestion and depletion remains a robust and straightforward method to remove the most easily and repeatedly detected peptides, clearing chromatographic, electrospray ionization and mass spectrometer space for improvements in identification coverage and quantification of low-abundance proteins. Our refined mechanistic analysis suggest that varying limited digestion times in combination with the use of other proteases with different site specificity constants (V/K) and different molecular-weight-cutoff filter sizes may hold the most potential to further improve coverage and quantitation of whole proteomes.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper (doi:10.1038/nmeth0314-345).

ACKNOWLEDGMENTS

This project was supported by the US National Center for Research Resources (5P41RR011823-17), National Institute of General Medical Sciences (8P41GM103533-17), National Institute of Diabetes and Digestive and Kidney Diseases (R01DK074798), National Heart, Lung, and Blood Institute (RFP-NHLBI-HV-10-5) and National Institute of Mental Health (R01MH067880). We thank D. Schwartz for help with motif alignments and J. Moresco, J. Savas and A. Pinto for comments on the manuscript.

AUTHOR CONTRIBUTIONS

These additional analyses and derivations were performed by B.R.F. and Mark S. Hixon, respectively. M.S.H. (Department of Biological Sciences, Takeda California, San Diego, California, USA) provided valuable assistance with describing the enzyme kinetics of these complex mixtures. B.D.S., K.J.W., T.X., J.C., S.K.P. and J.R.Y. agree with the reanalysis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ye, M., Pan, Y., Cheng, K. & Zou, H. *Nat. Methods* **11**, 221–222 (2014).
2. Liu, H., Sadygov, R.G. & Yates, J.R. III. *Anal. Chem.* **76**, 4193–4201 (2004).
3. Zybailov, B. *et al. J. Proteome Res.* **5**, 2339–2347 (2006).
4. Griffin, N.M. *et al. Nat. Biotechnol.* **28**, 83–89 (2010).
5. Schwanhäusser, B. *et al. Nature* **473**, 337–342 (2011).
6. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
7. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. *Nat. Methods* **6**, 786–787 (2009).

Erratum: Pouring over liquid handling

Vivien Marx
Nat. Methods **11, 33–38 (2014); published online 30 December 2013; corrected after print 13 January 2014**

In the version of this article initially published, the text stated that Toby Jenkins directs liquid handling for instrument manufacturer TTP Labtech. This name was misspelled; the correct name is Joby Jenkins. The error has been corrected in the HTML and PDF versions of the article.

Erratum: Dissecting genomic diversity, one cell at a time

Paul C Blainey & Stephen R Quake
Nat. Methods **11, 19–21 (2014); published online 30 December 2013; corrected after print 27 January 2014**

In the version of this article initially published, references 16–24 were incorrectly cited as references 15–23. The error has been corrected in the HTML and PDF versions of the article.

Corrigendum: Using networks to measure similarity between genes: association index selection

Juan I Fuxman Bass, Alos Diallo, Justin Nelson, Juan M Soto, Chad L Myers & Albertha J M Walhout
Nat. Methods **10, 1169–1176 (2013); published online 26 November 2013; corrected after print 27 January 2014**

In the version of this article initially published, the formula describing the connection specificity index (CSI) in Box 2 was incorrect. The denominator in the fraction of the CSI equation originally read "n_y"; the correct denominator is "# of X-type nodes in the network." The error has been corrected in the HTML and PDF versions of the article.

Corrigendum: Quantifying cell-generated mechanical forces within living embryonic tissues

Otger Campàs, Tadanori Mammoto, Sean Hasso, Ralph A Sperling, Daniel O'Connell, Ashley G Bischof, Richard Maas, David A Weitz, L Mahadevan & Donald E Ingber
Nat. Methods **11, 183–189 (2014); published online 8 December 2013; corrected after print 5 February 2014**

In the version of this article initially published, the current affiliation of author Ralph Sperling was not included. His current affiliation is the Fraunhofer ICT-IMM, Mainz, Germany. The error has been corrected in the HTML and PDF versions of the article.

Erratum: Singled out for sequencing

Kelly Rae Chi
Nat. Methods **11, 13–17 (2014); published online 30 December 2013; corrected after print 5 February 2014**

In the version of this article initially published, an incorrect institutional affiliation was given for Jie Wang: this affiliation was listed as Harvard when it should have been Peking University Cancer Hospital. The error has been corrected in the HTML and PDF versions of the article.

ONLINE METHODS

Software implementations. The software for fully automated generation of the simulated image data used in this study and the software for computation of the performance measures (**Supplementary Note 4**) were written in the Java programming language as plug-ins for the open bioimage informatics platform Icy[61] (**Supplementary Software**). Software implementations of the particle tracking methods of the participating teams (**Supplementary Note 1**) were written using various programming languages and platforms, including Java (stand-alone modules or plug-ins for ImageJ/Fiji[62] or Icy), C++ (provided as source code or executable) and Matlab (MathWorks).

Analysis of performance results. For each tracking method and each performance measure, 48 values could in principle be computed, corresponding to the 48 data cases (different combinations of particle dynamics, densities and signal levels). However, not all teams submitted tracking results for all cases, which ruled out the possibility to perform an overall comparison and ranking of the different methods based on all cases. We observed that teams who did not apply their method to all 48 cases generally focused on one or more of the four dynamics scenarios representing different biological applications, but even per scenario not all teams applied their method to all pertaining cases. Therefore, we decided to rank the methods according to best performance per measure and per data case (**Fig. 3** and **Supplementary Table 2**).

Verification of tracking results. Minor differences between the originally submitted tracking results and the verified results were to be expected because some of the software tools were converted to another platform to allow execution on the single evaluation system, and some methods were probabilistic in nature. Therefore, for each method, differences were considered acceptable (reproducible) if their means for each of α, β, JSC and JSC_θ were within 3% and the RMSE was within 0.5 pixel. In the vast majority of cases, the differences were acceptable, and the larger differences in some cases could be traced back to bug fixes and minor improvements in the software or parameter settings used for verification as compared to the original versions. In very few instances the results could not be verified owing to hardware or software limitations (**Supplementary Table 3**). For the analysis, the performance values computed from the originally submitted tracking results, not the verified results, were used.

Scoring of computation times. Computation times of all methods were measured on a single workstation (64-bit Intel Xeon X5550 2.67 GHz processor with 24 GB of RAM and running Microsoft Windows 7 Professional or Linux Fedora 16) to allow a fair comparison. We timed only those cases for which tracking results were submitted and verified. Similarly to the analysis of the accuracy performance measures, we ranked the methods according to best timing per data case (**Fig. 3** and **Supplementary Table 4**).

61. de Chaumont, F. *et al.* Icy: an open bioimage informatics platform for extended reproducible research. *Nat. Methods* **9**, 690–696 (2012).
62. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

doi:10.1038/nmeth.2808

ONLINE METHODS

Statistical information. The sample numbers for the experiments to compare ddPCR and canonical genotyping PCR (**Supplementary Fig. 1**) with the plasmid DNA were technical replicates. Those for the Surveyor assay and ddPCR experiments in HEK293T cells (**Supplementary Figs. 3** and **4**) were all biological replicates (independent transfections of HEK293T cells). We did not exclude any samples.

TALEN and CRISPR-Cas design and construction. TALEN Targeter (https://tale-nt.cac.cornell.edu/tutorials/talentargeterupdated)[20] was used to design the *PHOX2B* and the *PRKAG2* TALENs. ZiFiT (http://zifit.partners.org/ZiFiT/ChoiceMenu.aspx)[21,22] was used to design the *PKP2* and *BAG3* TALENs. TALENs were constructed using the Voytas laboratory's Golden Gate assembly system provided through Addgene (http://www.addgene.org/TALeffector/goldengateV2/)[23], except the backbone vector. We interchangeably used the Goldy[24] and the MR015 (gift from M. Porteus and M. Rahdar, Stanford University) TALEN backbones for the *PHOX2B* TALENs. We observed that these two backbones had similar activity in human iPS cells (data not shown). We used the *MR015* TALEN backbone for all other TALENs used in this study. The TALEN target sequences were aligned against the reference genome by BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to ensure that they were unique sites. We used the Asn-Asn (NN) repeat variable di-residues (RVDs) for all the TALENs used in this study. For *PHOX2B* and *PKP2* TALENs, the NN versions are described as the 'strong' TALENs. In addition, we made TALENs in which the binding sequence was the same, except that the NN RVDs were substituted with Asn-Lys (NK) RVDs (which induces less frequent recombination[25,26]) to create the 'weak' TALENs. For the individual TALEN target sequences, see **Supplementary Table 4**.

For the CRISPR-Cas system, we searched the reference human genome for unique CRISPR target sites ($GN_{20}GG$, where N is any nucleotide) that fit with the Church laboratory's system provided through Addgene (http://www.addgene.org/crispr/church/)[18], and found a unique site encompassing the RBM20 mutation site (GGTCTCGTAGTCCGGTGAGCCGG, where the underlined C was mutated to A). We followed the guidelines provided through Addgene to generate the gRNA expression construct.

Design of primers and probes for the TaqMan PCR system and oligonucleotide donors. To design the primers and probes, we used the TaqMan MGB Allelic Discrimination option in Primer Express 3.0 software (Life Technologies). For the *PHOX2B* mutation, the default setting (minimal amplicon size was 50 bp) was selected, whereas for the *PKP2*, *RBM20* and *PRKAG2* mutations, the amplicon size setting was changed to 100 bp to give optimal probe-primer sequences. For the *BAG3* mutation, we manually designed the primers and probe by following the instructions of Primer Express 3.0, as the default program did not give us any candidates. The point mutations were inputted as "SNPs for allelic discrimination," and designed custom TaqMan MGB probes were purchased from Life Technologies. Optimal primer sequences identified by the software were manually checked to ensure that at least one of the primers would anneal outside of the donor sequence. Oligonucleotide donors were manually designed to target the point mutation to the middle of the oligonucleotide sequence. Primers and oligonucleotide donors that underwent standard desalting purification were purchased from Integrated DNA Technologies. For sequences of the primers, probes and oligonucleotide donors, see **Supplementary Table 5**.

Culture and transfection of HEK293T cells. HEK293T cells were maintained in Dulbecco's modified Eagle medium with high glucose, sodium pyruvate and L-glutamine (Life Technologies) supplemented with 10% FetalPlex (Gemini Bio-Products) and 50 units/ml penicillin-streptomycin (Life Technologies). For transfection, 4×10^5 cells or 2×10^5 cells were plated into 12-well or 24-well plates, respectively. One day after plating, the cells were transfected with DNA using Lipofectamine 2000 (Life Technologies), according to manufacturer's instructions. For point mutagenesis by TALENs, 400 ng or 200 ng of each TALEN vector and 800 ng or 400 ng of an oligonucleotide donor DNA were transfected per well in 12-well or 24-well plates, respectively. For point mutagenesis by CRISPR-Cas, 400 ng or 200 ng of Cas9 vector, 400 ng or 200 ng of gRNA vector, and 800 ng or 400 ng of an oligonucleotide donor DNA were transfected per 12-well or 24-well plate, respectively. Genomic DNA was extracted from the cells 3 d after transfection with the DNeasy Blood & Tissue Kit (Qiagen).

Surveyor assay. The Surveyor assay was carried out according to manufacturer's instructions (Transgenomics). Using Lipofectamine 2000, 800 ng or 400 ng of each TALEN vector was transfected into HEK293T cells plated into a 12-well or 24-well plate, respectively. To PCR-amplify the target genomic regions, we used Phusion High-Fidelity DNA Polymerase (New England BioLabs) and the following primers: *PHOX2B* forward (FW), 5′-CTCCAGCCACCTTCTCCATA-3′; *PHOX2B* reverse (RV), 5′-CGCTGAGAAAGCTGAAGGTC-3′; *PKP2* FW, 5′-CAAAC TCAAGAATCTCATGATAACAGAA-3′; and *PKP2* RV, 5′-AC AAACCATCAAACAAACTGTG-3′. The digested DNA fragments were analyzed by 15% acrylamide gel electrophoresis. The intensities of the bands were quantified by ImageJ (http://rsbweb.nih.gov/ij/). Quantification of the frequency of NHEJ was calculated using the formula $100 \times (1 - (1 - a/(a + b))^{1/2})$, where a equals the integrated intensity of both cleavage product bands and b equals the integrated intensity of the uncleaved PCR product band.

Culture and transfection of iPS cells. For iPS cells, the University of California San Francisco Committee on human research #10-02521 approved the study protocol. The human iPS cell lines used in this study were generated from a healthy male patient, WTC10 and WTC11 (ref. 27), using the episomal reprogramming method[28]. We successfully applied our ddPCR and sib-selection method to another iPS cell line, WTB6, which was generated from a healthy female patient using the episomal reprogramming method[28] (data not shown). Informed consent was obtained for this procedure. iPS cells were maintained on Matrigel (BD Biosciences) in Essential 8 medium (Life Technologies), which was exchanged every other day. To sparsely populated wells (i.e., passaged wells), we added 10 μM Y-27632, a Rho-associated kinase (ROCK) inhibitor (Millipore), to promote cell survival[29].

For TALEN and donor vector transfections, we used the Human Stem Cell Nucleofector Kit-1 and a Nucleofector 2b Device (Lonza). WTC11 was used for the *BAG3* mutagenesis, whereas

doi:10.1038/nmeth.2840

WTC10 was used for the mutagenesis of other genes. When cells were 60–70% confluent in six-well plates, the medium was changed to Essential 8 with 10 μM Y-27632 at least 1 h before transfection. The medium was then transferred into a conical tube to be used as conditioned medium. Cells were washed with 2 ml/well phosphate-buffered saline (PBS) and treated with 500 μl/well Accutase (Millipore) at 37 °C for 5–10 min. Cells were then resuspended in PBS and counted using a Countess Automated Cell Counter (Life Technologies).

For each transfection, 2 million cells were then transferred into a conical tube and spun down at 1,500 r.p.m. for 5 min. The supernatant was aspirated, the cells were resuspended in 100 μl/transfection of Human Stem Cell Solution 1, and 3 μg of each TALEN vector (10 μg of each TALEN vector was used only in **Supplementary Fig. 6b**) and 6 μg of an oligonucleotide donor were added. The total volume of the DNA solution was less than 10 μl. The suspension was transferred to a cuvette and electroporated using program A-23. Conditioned medium (500 μl/transfection) was added to the transfected cells, which were then transferred to a conical tube with 9.5 ml of conditioned medium. The cells were plated into a Matrigel-coated 96-well plate using a multichannel micropipetter.

Detection of point mutations by ddPCR. The composition of the premixtures of allele-specific TaqMan probes and primers was 5 μM of an allele-specific FAM or VIC TaqMan MGB probe, 18 μM of a forward primer and 18 μM of a reverse primer in water. For optimization of the PCR annealing and extension temperatures, we mixed the following reagents in 1.5-ml tubes: 9 μl water, 12.5 μl 2× ddPCR Supermix for Probes (Bio-Rad), 1.25 μl FAM probe and primer premixture, 1.25 μl VIC probe and primer premixture and 1 μl 0.5 pg/μl 1:1 plasmid mixture of a wild-type allele and a mutant allele (25 μl total volume). Droplet generation with a QX100 Droplet Generator was performed according to the manufacturer's instructions (Bio-Rad), and the reaction was transferred into a 96-well PCR plate for standard PCR on a C1000 Thermal Cycler (Bio-Rad). The thermal cycling program conducted was: step 1, 95 °C 10 min; step 2, 94 °C 30 s; step 3, 50 °C–58 °C or 60 °C gradient 1 min; repeat steps 2 and 3 39 times; step 4, 98 °C 10 min. After the PCR was complete, the droplets were analyzed using a QX100 Droplet Reader (Bio-Rad) with the "absolute quantification" option. We first checked whether a probe set could properly discriminate between negative, FAM-positive, VIC-positive and double-positive populations. We then chose the highest annealing temperature that gave the best separation of the four distinct populations as the optimal temperature (**Supplementary Fig. 2**). The optimal temperatures differed between different probe-primer sets. For example, for the *PHOX2B* mutation shown in **Supplementary Figure 2**, the optimal annealing and extension temperature was 51 °C. For the *PKP2*, *RBM20*, *PRKAG2* and *BAG3* mutations, the optimal annealing temperatures were 57 °C, 57 °C, 51 °C and 55 °C, respectively (data not shown).

For ddPCR to detect point mutagenesis, we mixed the following reagents in 0.2 ml PCR 8-tube strips: 9 μl (for HEK293T cells) or 4 μl (for iPS cells) water, 12.5 μl 2× ddPCR Supermix for probes (Bio-Rad), 1.25 μl FAM probe and primer premixture, 1.25 μl VIC probe and primer premixture and 1 μl (for HEK293T cells, 100 ng) or 5 μl (for iPS cells, 50–150 ng) genomic DNA solution (25 μl total volume). Droplets were generated, which was followed by PCR at the optimal temperatures; the droplets were analyzed as described for the temperature optimization to calculated mutant allele frequencies in a wild-type allele background. For iPS cells, a multichannel micropipetter was used to directly transfer genomic DNA extracted from 96-well plates to reaction mixtures and then to transfer the reaction mixtures to a cartridge for generation of droplets.

Freezing iPS cells and genomic DNA extraction from iPS cells on a 96-well plate for sib-selection. When iPS cells (plated in 96-well plates) were 80–90% confluent, the cells were washed with 100 μl/well PBS and treated with 30 μl/well Accutase at 37 °C for 5–10 min. Half of the detached cells were placed into a 96-well plate for genomic DNA extraction, and the other half left in the original 96-well plate were frozen at −80 °C for future sib-selection.

For genomic DNA extraction, 50 μl of genomic DNA lysis buffer (10 mM Tris pH 7.5, 10 mM EDTA pH 8.0, 10 mM NaCl, 0.5% N-lauroylsarcosine and 1 mg/ml proteinase K (added fresh)) was aliquoted into each well on another 96-well plate. We transferred 15 μl/well of the cell suspension (in Accutase) to the 96-well plate containing the lysis buffer, and the mixture was combined using a multichannel micropipetter. The plate was then incubated at 55 °C overnight in a plastic container with a small amount of water (to prevent drying). Then, 100 μl 75 mM NaCl in ethanol that had been stored at −80 °C was added to each well, the solution was mixed and the plate was incubated at room temperature for 2 h. The solution was discarded by inverting the plate (the DNA remained adhered to the plate), and each well was washed with 100 μl 70% ethanol twice. The plate was then air-dried for 30–45 min to remove all residual alcohol. After drying, DNA in each well was dissolved in 30 μl TE buffer, and the DNA concentration was measured. For most wells, we obtained DNA concentrations of 10–30 ng/μl.

For cell freezing, 75 μl freezing medium (HyClone Fetal Bovine Serum (Thermo Scientific) with 10% dimethyl sulfoxide (Sigma)) was added to cells in each well of the 96-well plate and mixed by pipetting up and down, taking care to avoid creating too many bubbles. Then, 75 μl of mineral oil (Sigma, embryo culture grade) was layered onto the cell suspension. The 96-well plate was sealed with Parafilm and frozen at −80 °C in a Styrofoam container. When thawing the cells, the plate was placed in a 37 °C, 5% CO_2 incubator for 10–15 min. The cell suspension was transferred to a 1.5-ml tube, mixed with 0.5 ml Essential 8 medium with 10 μM Y-27632, and the number of cells was counted using a Countess Automated Cell Counter. An aliquot of the cell suspension was then removed and placed into a conical tube containing Essential 8 medium supplemented with 10 μM Y-27632 and mixed well. Cells were then plated into Matrigel-coated 96-well plates at the appropriate cell density.

For sib-selection, cell pools containing mutated cells were plated at a density of 500 cells/well into a 96-well plate, fed with Essential 8 medium with 10 μM Y-27632 every other day for 6 d and then fed with Essential 8 every other day for the next 4 d. After culturing cells in these conditions, they were 80–90% confluent and ready for the next round of sib-selection or cloning.

Isolation of iPS cell clones. The process of thawing a cell pool from which iPS clones were isolated was the same as described

for sib-selection (see above). For isolating iPS cell clones, the cells were plated at limiting dilutions for an average density of 100 cells/well into 96-well plates. The cells were fed with Essential 8 medium with Y-27632 every other day for 10 d, and then fed with Essential 8 every other day for the next 8 d or 10 d. By 18 or 20 d after plating, cells in only 10–30% of wells survived. Wells containing cells with pluripotent morphology that were likely to have been derived from a single clone were chosen for clonal analyses. We are aware that iPS cells purified via limiting dilution may be the progeny of more than one cell (mixed colony). However analysis via ddPCR allowed us to rapidly determine whether a putative clone is genetically homogenous by quantifying the alleles. After the initial purification we also frequently recloned cells to be sure that populations are homogenous (see below and **Supplementary Fig. 6g**). These wells were washed in 100 µl/well PBS, and cells were detached with 30 µl/well Accutase. Cells were divided into two even pools for genomic DNA extraction and freezing as described above, except that the transferred cells were strategically arranged in wells of new 96-well plates. In this way, the genomic DNA samples could be readily analyzed by genotyping PCR on a 384-well format, as described below. The concentration of genomic DNA extracted from the clones was 3–10 ng/µl, which was sufficient for genotyping PCR. Some investigators prefer cloning human iPS cells on feeder cells, and this process also works well with our method. However, we have found that using feeder-free culture conditions was more compatible with higher-throughput production and analysis in a 96-well plate format. However, it is essential to confirm homogeneity of the alleles via ddPCR and/or subsequent subcloning.

Genotyping PCR and expansion of isolated clones. The following reagents were mixed in each well of a 384-well plate: 2.5 µl 2× TaqMan Genotyping Master Mix (Life Technologies), 0.125 µl FAM probe and primer premixture (the same premixture used for ddPCR), 0.25 µl VIC probe and primer premixture (the same premixture used for ddPCR), 1.625 µl water, and either 0.5 µl extracted genomic DNA from iPS clones, 0.5 µl 0.5 pg/µl mutant allele plasmid/wild-type allele plasmid/1:1 mixture of both plasmids or 0.5 µl water (no-template control) (5 µl total). We made a master mix of these reagents excluding the templates, aliquoted 4.5 µl of the master mix into each well on a 384-well plate and added 0.5 µl of each template to the relevant wells. After mixing, the plate was centrifuged at 1,300 r.p.m. for 5 min. PCR genotyping was carried out using the Applied Biosystems 7500 or 7900 Fast Real-Time PCR System (Life Technologies) using the SDS2.4 software (Life Technologies). Plate documents of allele discrimination (AD) and the standard curve (AQ) for the samples were created according to the manufacturer's instructions. Preread was performed to measure background signals of each well with the AD plate document, a thermal cycling was performed with the AQ plate document, and finally postread and genotyping were done with the AD plate document (according to manufacturer's instructions). The thermal cycling parameters for the default condition of the AQ plate document were as follows: step 1, 50 °C 2 min; step 2, 95 °C 10 min; step 3, 95 °C 15 s; step 4, 60 °C 1 min; repeat steps 3 and 4 39 times. During thermal cycling, both the FAM and VIC signals were monitored.

The target genomic regions were amplified by conventional PCR with Phusion High-Fidelity DNA Polymerase from the genomic DNA of clones that were genotyped as positive for mutant alleles and sequenced to verify the mutations. The 96-well plates that had the positive clones were then thawed in the same way as for sib-selection (see above). The cell suspension was mixed with 0.5 ml Essential 8 medium with 10 µM Y-27632 in a 1.5-ml tube, and the cells were spun down at 6,000 r.p.m. for 5 min. Using a micropipetter, the surface oil layer followed by the medium was carefully removed. The cells were then resuspended in Essential 8 medium with 10 µM Y-27632 and plated into Matrigel-coated 6-well or 12-well plates. Wells containing iPS cells that partially differentiated during the cloning step were passaged at a 1:10 ratio 2–3 times, to remove these cells and recover pools that are fully pluripotent.

We confirmed that the mutant allele:wild-type allele ratio was 1:1 and 1:0 for heterozygous and homozygous mutant lines, respectively, by ddPCR to verify that they were pure clones. The results for the *PRKAG2* mutant heterozygote are shown in **Supplementary Figure 6g** as an example.

Fluorescent staining of iPS cells. iPS cells cultured in 24-well plates were washed with 500 µl/well PBS, fixed with 250 µl/well 4% paraformaldehyde at room temperature for 20 min and then washed with 500 µl/well PBS three times. The cells were blocked and permeabilized with 250 µl/well PBS with 0.1% Triton-X (PBS-T) supplemented with 5% BSA at room temperature for 1 h, and then in 200 µl/well primary antibodies diluted 200 times in PBS-T with 5% BSA at 4 °C overnight. The primary antibodies (all purchased from Abcam) used in this study were SOX2 (ab59776), OCT4 (ab19857), SSEA4 (ab16287) and TRA-1-81 (ab16289). The cells were washed with 500 µl/well PBS-T three times, incubated with 200 µl/well secondary antibodies diluted 500 times in PBS-T with 5% BSA at room temperature for 1 h and then washed with 500 µl/well PBS three times. The secondary antibodies used were Alexa Fluor 488–labeled goat anti-mouse IgG (A-11001) and Alexa Fluor 488–labeled (A-11008) or Alexa Fluor 594–labeled (A-11012) goat anti-rabbit IgG (Life Technologies). Finally, the cells were mounted with 75 µl/well VectaShield Mounting Medium with DAPI (H-1200) (Vector Laboratories). The pictures were taken by using BZ-9000 microscope (Keyence).

Karyotyping. Karyotyping was carried out by Cell Line Genetics.

Sequencing of potential off-target genomic sites. Potential off-target sites for the *PHOX2B* and *PRKAG2* TALENs were predicted by a recently reported bioinformatic tool, TALENoffer[30]. We set the distance between two TALEN binding sites from 12 bp to 24 bp, and other parameters were the default settings for the prediction. For both TALEN pairs, the top candidates were their on-target sites, indicating that the prediction was properly done. We chose top 10 off-target sites for each TALEN pair and designed primers that amplify 400–1,100-bp genomic regions around the off-target sites. The genomic regions were amplified by using Phusion High-Fidelity DNA Polymerase from the *PHOX2B* mutant heterozygous, *PHOX2B* mutant homozygous, *PRKAG2* mutant heterozygous, PRKAG2-mutant heterozygous 2 and independent *PRKAG2* mutant heterozygous lines as well as the parental iPS cell line, WTC10. The amplicons were sequenced from both 5′ and 3′ ends by using the primers that were used to amplify them.

doi:10.1038/nmeth.2840

All the sequencing results from the mutant lines were compared to those for WTC10. We did not see any mutations in the mutant lines (**Supplementary Table 3**).

20. Doyle, E.L. *et al. Nucleic Acids Res.* **40**, W117–W122 (2012).
21. Sander, J.D. *et al. Nucleic Acids Res.* **38**, W462–W468 (2010).
22. Sander, J.D., Zaback, P., Joung, J.K., Voytas, D.F. & Dobbs, D. *Nucleic Acids Res.* **35**, W599–W605 (2007).
23. Cermak, T. *et al. Nucleic Acids Res.* **39**, e82 (2011).
24. Bedell, V.M. *et al. Nature* **491**, 114–118 (2012).
25. Christian, M.L. *et al. PLoS ONE* **7**, e45383 (2012).
26. Cong, L., Zhou, R., Kuo, Y.C., Cunniff, M. & Zhang, F. *Nature Commun.* **3**, 968 (2012).
27. Kreitzer, F.R. *et al. Am. Journal Stem Cells* **2**, 119–131 (2013).
28. Okita, K. *et al. Nat. Methods* **8**, 409–412 (2011).
29. Watanabe, K. *et al. Nat. Biotechnol.* **25**, 681–686 (2007).
30. Grau, J., Boch, J. & Posch, S. *Bioinformatics* **29**, 2931–2932 (2013).

doi:10.1038/nmeth.2840

ONLINE METHODS

Annotation sources. We acquired a range of annotations at different scales and in a variety of data formats. Here we group data types by class and name their sources.

Open chromatin. We used DNase I hypersensitivity assay followed by sequencing (DNase-seq) and formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq) peak calls and DNase footprints from ENCODE.

Transcription factor binding. We used ChIP-seq peak calls for 124 transcription factors from ENCODE, JASPAR[15] motifs aligned under corresponding factor chromatin immunoprecipitation–sequencing (ChIP-seq) peaks from Ensembl and bound transcription factor binding motif data from ENCODE.

Histone modifications. We used ChIP-seq peak calls for 12 different modifications from ENCODE.

RNA polymerase binding. We used ChIP-seq peak calls from ENCODE.

CpG islands. We used predictions from Ensembl[14].

Genome segmentation. We used Ensembl integration[16] of the ENCODE SegWay[17] and ChromHMM[18] segmentation calls, which identified seven discrete states.

Conservation. We used genomic evolutionary rate profiling (GERP) scores from mammalian alignments from the Sidow laboratory at Stanford University, both at the specific variant nucleotide and averaging over 100 base pairs surrounding each variant[19].

Human variation. We took variants, allele frequencies and ancestral allele calls from 1000 Genomes Project phase 1 data. Mean heterozygosity and mean derived allele frequency of variants were both calculated in 1 kb windows from global population frequencies.

Genic context. We used distance to the nearest TSS from Gencode annotation provided by ENCODE[20], and distance to the nearest splice site and summary gene region annotations (any base annotated as exonic, intronic, coding sequence, 5′ or 3′ untranslated region, splice site, or start or stop codon in any transcript) from Gencode annotation provided by Ensembl.

Sequence context. We used sequence context information from the GRCh37 assembly of the human genome produced by the Genome Reference Consortium[21] including G+C content calculated over the 100 bp surrounding each variant, a Boolean variable indicating whether the variant is in a CpG context in the reference assembly, reference nucleotide at the variant position. We also used a Boolean variable indicating if the variant falls in repeat sequence from the University of California at Santa Cruz genome browser.

We developed a pipeline that can apply these annotations to a given set of variant loci. The result was a large matrix with a row for each variant locus and a column for each possible annotation. The column type, which depends on the annotation class, can be: (i) the number of cell lines in which the variant locus overlaps some annotation, such as DNase I hypersensitive sites and ChIP-seq peaks, (ii) a present-absent binary flag for the annotation at the variant locus (for example, whether this region is ever in an annotated intron) or (iii) a continuous value for genome-wide annotations, such as conservation and distance to the nearest TSS.

Construction of disease and control variant sets. The disease-implicated set of variants was composed of all variants annotated as 'regulatory mutations' from the April 2012 release of HGMD and downloaded from Ensembl release 70. After we removed variants that has the same position, a set of 1,614 disease-implicated SNVs remained. For all three control sets, we used variants identified in the low-depth whole-genome study in the 1000 Genomes Project (1KG) phase 1 release, downloaded from the project website in December 2012. We limited our analysis to variants with minor allele frequency ≥1% to reduce the chance of including rare functional variants in our control set. (We performed sensitivity analyses by either focusing exclusively on variants from European populations or on rare, singleton variants, and we found qualitatively similar results for cross-validation with the common variant controls; data not shown.)

As we only had SNVs in our 'disease' set, we also limited our analysis to SNVs in our control sets, for a total of 15,730,276 potential control SNVs. The first control set was a random selection of SNVs from across the genome, 100 times the size of the disease set, to get a reasonable sample of the background while making analyses computationally tractable. The second control set included 1KG SNVs matched for distance to the nearest TSS genome-wide, but not necessarily near the same genes as the HGMD variants. This set was 10 times the size of the disease set; for larger sets we could not ensure that the distributions of distances matched those of the HGMD variants. The final control set was composed of all 1KG variants in the 1 kb surrounding each of the HGMD variants (n = 5,027 variants).

Individual feature analysis. We investigated whether any of the annotations showed a different distribution in the disease and control sets. Annotations can be grouped into two classes: a large class of regional data indicating whether or not each variant lies in an annotated element, possibly across multiple cell lines, and a smaller class consisting of several continuous variables.

For the regional data, we ignored the number of cell lines in which a variant was found (as these were not independent across cell lines) and just used a single binary variable per feature indicating whether each variant was found in this element in any cell line. Annotations not specific to a cell line are already binary. For each feature we then computed a contingency table identifying how these counts differed in our disease and control sets. We used Fisher's exact test to compute the significance of enrichment or depletion.

For continuous features, we used a two-sided Mann-Whitney U test to establish whether there was a significant difference in the distribution of each feature between the two classes. We used this test (here and throughout this study), as it does not make any assumptions about the underlying distributions of our data. For the measures of the distance to the nearest TSS or splice site, we used absolute values, though we supplied the original signed value to the classifier as it may be informative to take into account whether the variant is upstream or downstream from the nearest TSS. All P values were adjusted using the Bonferroni correction for multiple testing. Unadjusted P values are also reported (**Supplementary Table 5**).

Classifier algorithm. Our classifier needs to simultaneously handle a large number of continuous and categorical features. Two of the control sets are also very unbalanced with respect to variant class, in that there are considerably fewer disease-implicated

variants than controls. To address these issues, we used a slightly modified version of the random forest algorithm[10]. Random forests are a robust and widely used approach to classification that can deal with the different feature types that we use. They are also robust to the presence of features that are not predictive (so we did not perform any feature selection). We modified the standard algorithm to address class imbalance by sampling equally from both classes when generating the training set for each component decision tree in the forest. The even class distribution means that each tree is trained on a smaller subset of control variants, but we used enough trees that most of the controls should be used at least once in the full model (subject to the normal random subsampling that is part of the algorithm). The random forest approach also has the advantage that it allowed us to compute the relative importance of each feature from the trained model.

We trained three forests, one for each of the three control variant sets and using the same disease variants as the positive set in each forest. We experimented with different numbers of decision trees in the forest and found that performance seemed to saturate around 100 trees, and this number should also ensure that we sample a good proportion of variants in each of the training sets. We used the mean AUC value across each of the test sets in each fold as our main measure of classifier performance.

One potentially confounding characteristic of the HGMD data is that some genes have multiple associated variants (mean = 2.03, median = 1), some of which are located physically close and may have annotations in common. When performing cross-validation, variants from the same gene that appear in both the training and test sets may inflate performance statistics. To control for this, we created a stringent set of disease variants in which a single variant is randomly selected for each gene, and we observed a similar performance pattern, with slightly reduced AUC values (0.95, 0.82 and 0.64, respectively).

All software was written in the Python language, using a random forest implementation from the Scikit-learn library[22]. The modified source code is available at http://www.sanger.ac.uk/resources/software/gwava/ and as **Supplementary Software**.

Feature importance. To identify which features contribute to the discriminative ability of each classifier, we computed the relative Gini importance of each feature across each component tree of the three forests (**Supplementary Fig. 3**). Gini importance measures the mean decrease in impurity at each node in the tree owing to the feature of interest, weighted by the proportion of samples reaching that node.

Classifier score distribution. We computed the distribution of scores across all variants from the 1000 Genomes Project on chromosome 16 (with variants included in any training set removed; **Supplementary Fig. 8**). Although the distributions were somewhat different for each classifier, as expected, few variants were assigned high scores by any version. These distributions allowed us to compare scores from any candidate variant with the background distribution to estimate how 'unexpected' any given score is.

Validation experiments. *Annotating pathogenic variants from ClinVar.* We downloaded the full ClinVar database in VCF format in early 2013 (file name: clinvar_20130118.vcf), identified all variants annotated as 'pathogenic' (those with US National Center for Biotechnology Information (NCBI) clinical significance code = 5 in the "INFO" field) and extracted them. We first removed all variants that overlapped any coding sequence or essential splice sites (as annotated in Ensembl release 70) and then any variants overlapping with an HGMD variant. The resulting set of 194 variants constitutes the set of pathogenic noncoding variants we used in this analysis. We performed a similar filtering to identify all likely nonpathogenic variants annotated (those with NCBI clinical significance code 2 or code 3) and derived a set of 150 nonpathogenic noncoding variants. We also constructed a control set matched for distance to the nearest TSS from the 1000 Genomes Project data as described above for the HGMD variants, and again we only included 1000 Genomes variants with mean allele frequency ≥1%, and we included 100 control variants for each ClinVar variant, which resulted in a set of 19,400 control variants. We annotated these three sets of variants with the classifier trained on variants matched by distance to the nearest TSS and compared the classification results with ROC curves (**Supplementary Fig. 4**).

Annotating GWAS SNPs. We downloaded the GWAS catalog from the US National Human Genome Research Institute website in December 2012 and identified all variants with a "Context" field implying the variant did not fall in coding sequence. For the matched control set, we used a list of SNVs from common GWAS genotyping arrays constructed using information from Ensembl release 70, and overlapping with variants from the 1000 genomes project. We selected ten matching SNVs for each GWAS signal. The genotyping platforms used were Affymetrix GeneChip 100K, GeneChip 500K and SNP6, and Illumina HumanCNV370 Quadv3, HumanHap300v2, HumanHap550v3.0, Cardio Metabo, Human1M-duoV3 and Human660W-quad.

We compared the score distributions of these two sets of variants with a two-sided Mann-Whitney *U* test (**Supplementary Fig. 5**).

We downloaded the replication status annotations available in the supplementary material of ref. 11. We used these annotations to stratify the classifier scores according to whether the annotated SNPs were not validated, were internally validated or were validated in an independent study (**Supplementary Fig. 6**). Comparison of score distributions was performed with a two-sided Mann-Whitney *U* test. The *P* values comparing all pairwise combinations of these three sets of variants are: not replicated versus internally replicated, $P = 2.56 \times 10^{-9}$; not replicated versus independently replicated: $P = 3.65 \times 10^{-7}$ and internally replicated versus independently replicated, $P = 0.024$.

Application to personal genomics. We downloaded variant calls for the individual NA06984 from the 1000 Genomes Project website, and identified all variants found on chromosome 22 in this individual. We created a training set for the classifier based on the control set matched for distance to the nearest TSS but with all variants on chromosome 22 removed. We then built a classifier using the same approach described earlier on this reduced training set. We used this classifier to annotate all variants from the NA06984 chromosome 22 and the 33 HGMD variants from the same chromosome, and used a ROC curve to demonstrate how well we can discriminate the HGMD variants from background (**Supplementary Fig. 7**).

For individual gene analysis, we used the 24 unique genes annotated in HGMD as being affected by this set of 33 variants. For genes associated with more than one variant, we randomly

doi:10.1038/nmeth.2832

selected a single variant and disregarded the rest. We downloaded the coordinates from each of these genes from Ensembl and identified all variants from NA06984 that overlapped the gene region ± 5 kb (the distance used by Ensembl to associate a variant with a gene). We removed any variant overlapping coding sequence or an essential splice site. For each gene, we then computed the GWAVA score using the classifier trained on the control set matched for distance to the nearest TSS and identified the rank of the HGMD variant at each locus (**Supplementary Table 4**). To test the significance of this result, we developed some simulation software (available at the FTP site above, along with all other software) to establish how often we would expect to find a result as extreme as or more extreme than that observed if we were ranking the variants around each gene at random. We used this software to derive empirical *P* values for our results based on 1,000,000 random samples.

Application to somatic mutations. We downloaded all annotated noncoding somatic mutations from the COSMIC database, release 64, in March 2013 and limited our analysis to those annotated as being discovered in a whole-genome study. We identified all mutation loci that are found in more than one study (according to the COSMIC study identifier) and annotated these as recurrent. Comparison of score distributions was performed with a two-sided Mann-Whitney *U* test (**Fig. 2**).

Comparison with MutationTaster. We uploaded all noncoding somatic mutations from whole-genome studies in COSMIC release 64 that did not overlap either coding sequence or essential splice sites to the MutationTaster webserver in October 2013, and we obtained predictions for 93,692 unique mutations that could be mapped to a transcript model. MutationTaster reports multiple predictions for mutations that overlap multiple transcripts, and we computed a unique prediction for each mutation by assigning the prediction "disease_causing" to any mutation with this prediction in any transcript and "polymorphism" otherwise. We discarded variants with a prediction of "polymorphism_automatic" as these are made by database lookup (n = 1,340 variants). We used contingency tables to compare the number of variants predicted as "disease_causing" with whether or not the mutation was recurrent in different studies, and used Fisher's exact test to compute the significance of the enrichment. To compare this result with that of GWAVA, we assigned GWAVA scores to the same 92,352 mutations and threshold the GWAVA score with mutations scoring >0.5 identified as "functional" and all others "nonfunctional," and again used a contingency table to compute the enrichment of recurrent mutations among those called as functional.

Classifier availability. The GWAVA web server allows users to retrieve precomputed scores from each of the three classifiers for all known germ-line and somatic SNVs found in Ensembl release 70. All the underlying annotations used by the classifier are also available at http://www.sanger.ac.uk/resources/software/gwava/.

The source code, documentation, set of annotations used, all variant data sets described here and a plugin for the Ensembl Variant Effect Predictor[23] are available from the FTP server linked from the GWAVA webpage.

15. Mathelier, A. *et al. Nucleic Acids Res.* **42**, D142–D147 (2014).
16. Hoffman, M.M. *et al. Nucleic Acids Res.* **41**, 827–841 (2013).
17. Hoffman, M.M. *et al. Nat. Methods* **9**, 473–476 (2013).
18. Ernst, J. & Kellis, M. *Nat. Methods* **9**, 215–216 (2012).
19. Davydov, E.V. *et al. PLoS Comput. Biol.* **6**, e1001025 (2010).
20. Harrow, J. *et al. Genome Res.* **22**, 1760–1774 (2012).
21. Church, D. *et al. PLoS Biol.* **9**, e1001091 (2011).
22. Pedregosa, F. *et al. J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
23. McLaren, W.M. *et al. Bioinformatics* **26**, 2069–2070 (2010).

doi:10.1038/nmeth.2832

ONLINE METHODS

Device fabrication. Nanofluidic devices were fabricated out of PDMS (1:10 cross-linker:base, Sylgard 184, Dow Corning) using standard soft lithography and molding techniques. Briefly, two master molds were fabricated out of SU-8 photoresist (MicroChem) on 4-inch silicon wafers: one for the flow layer and one for the control layer. The flow-layer master mold had two different thicknesses of features of 1 and 5 μm. The control layer features were 15 μm thick. To make the PDMS devices, we poured a 4-mm-thick layer of PDMS onto the control-layer master mold and baked at 80 °C for 1 h to cure. The PDMS was then peeled off the mold, cut into individual chips, and punched with inlet and outlet holes. Concurrently, a 10-μm-thick layer of PDMS was spin coated onto the flow-layer mold and baked at 80 °C for 3 h. The PDMS-coated flow-layer mold and control-layer pieces were oxygen-plasma treated, aligned using fabricated alignment marks, and pressed together to create a permanent bond. The PDMS devices were then oxygen-plasma bonded to clean coverslips. This fabrication procedure resulted in the SWIFT devices (**Fig. 1b**), with a 10-μm membrane between the measurement channels and the control channel. Once devices were bonded to coverglasses, the PDMS sides of the channels and the glass coverslip were covalently coated with PEG simultaneously following the protocol in ref. 7.

Sample preparation. *Nup153FG.* The nucleoporin 153 (Nup153FG amino acids 875–1475 of the full-length Nup153) contained a single cysteine mutation at position 917. The protein was expressed and purified as described previously[21]. Briefly, Nup153FG containing a C-terminal chitin-binding domain followed by an intein, and an N-terminal His tag followed by a TEV protease cleavage site, was expressed from a pTXB3 vector in *Escherichia coli* BL21(AI) in terrific broth (TB) medium at 37 °C. All purification buffers contained 0.2 mM tris(2-carboxyethyl)phosphine (TCEP) and 1 mM phenylmethanesulfonyl fluoride (PMSF). The measurement buffer was 1× PBS, pH 7.4, 10% glycerol, 0.2 mM TCEP and 1.25 mM Trolox.

170- and 201-bp DNA. Forward primer 5′-GGGCTTATGTG ATGGACCC-3′ and labeled reverse primer 5′-(Alexa 647)-GAGTCGCTGTTCAATACATG-3′ were used to amplify a region of the pUC19 plasmid that contains a 201-bp Widom 601 nucleosome positioning sequence[22] for 201-bp dsDNA. Forward primer 5′-(Alexa 594)-CCCTATACGCGGCCGCCCTGGAGAATCCC GGTGCCGAGGCCGCT-(Alexa 488)-CAATTG-3′ and reverse primer 5′-CATGCACAGGATGTATATATCTGACACGTGCCT-(Alexa 594)-GGAGA-3′ were used for the 170-bp dsDNA. PCR was carried out using Phusion high-fidelity DNA polymerase (NEB). PCR product was then purified by extracting DNA from an agarose gel using a gel extraction kit (Invitrogen). The measurement buffer for 201-bp dsDNA was 1× PBS, pH 7.4, and 1.25 mM Trolox.

2,400-bp DNA. A random sequence of 2,400 bp from pQE60 plasmid was cloned in pUC19 plasmid and amplified using the following dual-labeled forward primer and reverse primer to generate a FRET standard: forward primer: 5′-(Alexa 594)-GGGCTTATGTGA-(Alexa 488)-TGGACCC-3′ and reverse primer: 5′-GGGCTTATGTGATGGACCC-3′. PCR and purification of amplified DNA was done similarly to that for the 201-bp DNA and followed by an additional size-exclusion step. The measurement buffer was 1× PBS, pH 7.4, and 1.25 mM Trolox. All labeled primers were purchased from IBA.

Nucleosomes. Nucleosomes were reconstituted *in vitro* from *Drosophila melanogaster* histones and a DNA fragment of 170-bp DNA following well-established protocols[14,23]. Briefly, the 170-bp fragment was amplified via PCR as described above. Core histones (H2A, H2B, H3, and H4) without any additional tag were expressed recombinantly in *E. coli* BL21(DE3) pLysS cells and purified from inclusion bodies by cation-exchange chromatography. Histone concentration was quantified by BCA assay (Thermo Fisher). Histone octamers were prepared by mixing the four histones under denaturing conditions (7 M guanidinium hydrochloride) and subsequent dialysis against a high ionic strength buffer (2 M NaCl). Octamers were purified on a Superdex 200 gel filtration column (GE Healthcare) and monitored using SDS-PAGE. Fractions showing equimolar amounts of histones were added to 25% glycerol and kept at −20 °C for long-term storage. Finally, nucleosomes were assembled by initially mixing the 170-bp DNA with histone octamers under high salt concentration (2 M NaCl) followed by stepwise salt dialysis to 10 mM Tris, 0.5 mM EDTA buffer at pH 7.5. Nucleosome assembly was verified by native electrophoresis on an agarose gel. The octamer-to-DNA mass ratio was empirically adjusted (usually between 0.7 and 1.3) so that only a single band of lower electrophoretic mobility than that of free DNA was observed after the gel was stained with ethidium bromide. BSA at 0.1 mg/mL was added to all buffers in order to minimize nucleosome dissociation upon dilution. The measurement buffer was 10 mM Tris, pH 7.5, 0.5 mM EDTA, 25% glycerol, and 1.25 mM Trolox.

Holliday junction. The following four DNA strands were purchased from IBA to assemble Holliday junctions: H strand, 5′-(Alexa 488)-CCGTAGCAGCGCGAGCGGTGGGCGAACGCTTA-3′; B strand, 5′-(Alexa 594)-CCCTAGCAAGCCGCTGCTACGG-3′; X strand, 5′-GGGCGGCGACCTACGAAACGG TCCCAGTTGA GCGCTTGCTAGGG-3′; and R strand, 5′-TAAGCGTTCGCCC ACCGCTCGGCTCAACTGGGACCGTTTCGT-3′. The four strands were annealed at equal concentration of 10 μM by slowly cooling from 90 °C to 4 °C over 2 h in a thermocycler in 10 mM Tris, pH 7.5, 50 mM NaCl, 1 mM EDTA buffer. The measurement buffer was 10 mM Tris, pH 7.5, 50 mM NaCl, 10 mM or 50 mM $MgCl_2$, 25% glycerol and 1.25 mM Trolox.

Transglutaminase 2. Following ref. 24, human TG2 was cloned in pBAD vector flanked at N′ and C′ termini with Snap-tag and Clip-tag (NEB), respectively, in an intein-12His and GST-tag fusion construct and expressed in Bl21(AI) for 16 h at 30 °C without any induction[24]. TG2 was purified under native conditions following standard purification protocols for His-tagged proteins. Intein tag and GST tag were cleaved from C′ termini by cleaving with TEV protease and exchanged to 10 mM Tris, pH 7, 0.2 mM TCEP, and 1 mM EDTA. Labeling with an equal concentration of Snap-tag Alexa 594 and Clip-tag Alexa 488 (Invitrogen) dye was performed at room temperature for 1.5 h, and excess dye was removed through gel filtration column purification. The measurement buffer was 10 mM Tris, pH 7 or pH 4, 200 mM NaCl, 0.2 mM TCEP, 25% glycerol, and 1.25 mM Trolox.

SWIFT microscope setup. For SWIFT microscopy, we used a custom-built TIRF microscope centered around a 100× high–numerical aperture (NA) total-internal-reflection fluorescence

(TIRF) objective (UAPON 100XOTIRF, NA 1.49, Olympus). Three lasers were available for excitation: a LuxX 488 nm 200 mW (Omicron) with a 475/23 laser cleanup filter (Semrock), a sapphire 568-nm 100-mW laser (Coherent), and an OBIS LX 660-nm 100-mW with a 661/11 laser cleanup filter (Semrock). The excitation from the lasers was routed through 10× beam expanders (Qioptic). The lasers were combined with dichroic mirrors (Semrock), concentrically aligned, and then routed into the back port of the microscope by a mirror on a laterally adjustable stage (M505, PI) that was used to control the TIRF angle. An adjustable rectangular aperture was also installed in the excitation path.

A custom 488/568/660 trichroic (AHF, pass bands 500–550; 590–655; 665–740) was used to split the excitation and emission light. The collected fluorescence was filtered through a double notch (Stopline 568/647, Semrock) and split by a dichroic mirror (alternatively, by a 561-nm and a 643-nm dichroic (Chroma)) between the microscope and the camera to separate the emitted light into two spectral bands on two camera chips for FRET experiments. 525/50 and 617/73 emission filters were used, respectively. An iXon X3 EM+860 camera and an iXon Ultra 897 (Andor Technology) with 512 × 512–pixel chips were used to collect the emitted photons.

The SNR given in the text (see also **Supplementary Fig. 4** and **Supplementary Note 3**)[12,25] was calculated as $SNR = (I_s - I_b)/\sigma_b$, where I_s is the mean signal intensity of 1,000 traces (intensity for each point was calculated as the total intensity measured in a 16 × 16–pixel area around the particle), I_b and σ_b are, respectively, the mean and s.d. of the background intensity of the same traces (where the background intensity for each point is calculated as the mean value along the entire movie of the summed intensity of the 16 × 16–pixel area around the particle).

FRET experiments were performed by exciting only the donor molecule but recording both the donor and acceptor emissions. The cross-talk between the channels was characterized using donor-only samples, and it was determined that 15% of the Alexa 488 donor signal leaks into the Alexa 594 acceptor channel. Further, the donor (D) and acceptor (A) could be excited alternatively to verify the presence and viability of the acceptor, following the general idea of the alternating-laser excitation concept (ALEX)[13].

Supplementary Video 9 shows DNA labeled with Alexa 647 taken on a commercially available Olympus Cell^R TIRF microscope using a 640-nm laser, a Hamamatsu Image EM CCD camera, a 60× oil-immersion objective, and quadband dichroic and emission filters (EMBL Advanced Light Microscopy Facility).

Device operation. A schematic of the microfluidic device is shown in **Figure 1b**. Solutions were stored in 0.5-mL microcentrifuge tubes that were interfaced to the device inlets and outlet via #30 Tygon tubing (details of this configuration can be found in ref. 10; **Supplementary Fig. 1** shows the SWIFT device mounted on the TIRF setup and the working principle). Sample and buffer solution are fed into the device and combined in a standard laminar sheath-flow mixer. The buffer and sample are mixed in a ratio of 100:1 at the normal operating pressure of 0.54 p.s.i. on the buffer inlet and 0.32 p.s.i. on the sample inlet, which we apply with hydrostatic pressure on the microcentrifuge tube reservoirs. The concentration of the sample molecules is homogenized across the width, w, of the channel within time $t = w^2/8D$ for a sample with diffusivity D (the working principle and the dead time of the mixer are explained in **Supplementary Fig. 3**). In order to avoid crowding, which can make measuring traces and automated tracking difficult, we used concentrations of ~100 pM.

The control channel is attached to a source of pressurized nitrogen (between 29 and 36 p.s.i.) via a 22-gauge Luer stub and Tygon tubing that is controlled with a pressure regulator (Porter Instruments) and measured with a digital gauge. Two outlet channels after the mixing region are designed with lower resistance to maintain the flow of sample and buffer through the mixer and to prevent undesired diffusional mixing before they enter the mixing channel. The average time molecules spend in the measurement region is controlled by adjusting the difference between the elevation of microcentrifuge tubes connected to the buffer and sample outlet (**Supplementary Fig. 1**).

In order to increase the photostability of the fluorophores, we used a similar approach to that described in ref. 10: specifically, solution deoxygenation to alleviate bleaching and the addition of a reducing and oxidizing system (ROXS) to ameliorate blinking[26]. The deoxygenation was performed by pressurizing the control channel with nitrogen. Because PDMS is permeable to gases, the nitrogen from the control valve diffuses through the PDMS into the measurement channel, where it displaces the oxygen originally present in the sample solution. The flow rate of nitrogen from the control channel through the PDMS was 0.09 µL/min at 36 p.s.i. control-channel pressure. The simulation in **Supplementary Figure 4** and **Supplementary Video 3** was performed using Comsol Multiphysics. Additional nitrogen channels surrounding the control channel and inlets were constantly ventilated with nitrogen at 1 p.s.i. to further ensure deoxygenation.

A triplet quencher can be added either to the sample using the on-device mixer or to the sample buffer before loading into the device. Trolox at a total concentration of 1.25 mM (300 µM converted to Trolox quinone via exposure to UV light; concentration was determined with $A_{255\,nm}$) was used as an efficient triplet quencher and prepared as previously described in ref. 26.

Channel depth measurement. The depth of the measurement channels on the mold for the flow layer was measured using a profilometer (Dektak), which corresponds to the depth of channels in device with 0 p.s.i. pressure in the control channel ($d_{0\,psi}$). The SWIFT device was filled with 0.1 µM FITC dye solution. The fluorescence micrographs of the channels (**Fig. 1c**) were obtained on an inverted fluorescence microscope (Zeiss Axiovert) with an HBO 100 lamp and a GFP filter set at various pressures of N_2 in the control channel. Pressurization of the control channel with gas reduces the depth of the underlying measurement channels and, hence, the volume of the measurement channels. It was ensured that the experiments were performed in the linear detection regime of the detectors so that fluorescence intensity was directly proportional to the volume of the measurement channels. The intensity in the 0-p.s.i. micrograph ($I_{0\,psi}$) was taken equal to the depth of channels in the flow-layer mold ($d_{0\,psi}$). Other intensity values (I) in micrographs were converted to depth (d) using the formula $d = I \times d_{0\,psi}/I_{0\,psi}$.

doi:10.1038/nmeth.2809

SWIFT microscopy tracking program. A custom software program written in Igor (WaveMetrics) was used to analyze the videos of individual molecules. Molecules were localized in each frame using the "Localizer package" (P. Dedecker, Katholieke University, Louvain, Belgium) for IgorPro by applying a threshold on the generalized likelihood-ratio test. Trajectories were then constructed by connecting the molecules localized in different frames with the following algorithm.

1. Particles detected in the first frame were considered the starting points of new trajectories.
2. For all successive frames, particles detected were connected to previous trajectories if detected within a maximum distance of 25 pixels and two frames. Whenever an ambiguity occurred (when a particle was assignable to multiple trajectories), the particle was not assigned to any of the potential trajectories. (Such a method is highly inefficient but guarantees the highest reliability.) If a particle was not connected (or potentially connectable) to any previously detected particle, then it was considered the start of a new trajectory.
3. From the found trajectories, the average jump distance was used to estimate the flow rate.
4. Different trajectories were then further connected using the knowledge of the flow rate to estimate the position of detected particles along the flow axis. Whenever an ambiguity occurred, the trajectories were not connected.
5. For alternating excitation of D and A dyes, the two cameras imaging the different wavelengths were aligned on the basis of fiducial beads (TetraSpeck, Invitrogen). The tracking was done on the sum of donor and acceptor channel frames recorded upon D excitation (t). The intensity for the same particle upon A excitation ($t + 1$) was calculated by extrapolating the position of the particle between two successive D excitation frames (t and $t + 2$) as the position corresponding to the maximum intensity value in the pixel window centered around the average position between the frames t and $t + 2$.

Diffusion coefficient analysis. Unlike in traditional single-molecule TIRF experiments, molecules in the SWIFT device are not immobilized on the surface but instead diffuse and flow through the channels. This affects the SNR because the photons emitted by the fluorophore can be distributed over multiple camera pixels during a single exposure and averaged with more background. For this reason and for automated particle tracking, large molecules with low diffusion coefficients are more suitable for this technique than small, quickly diffusing ones. The SNR is also affected by diffusion of the molecule axially between regions of varying excitation (owing to the exponentially decaying intensity of the TIR field).

The motion of the molecules is a convolution of diffusion and flow, both of which can be obtained by fitting the mean-squared displacement (MSD), $\rho(t) = [r(t) - r(0)]^2$, where $r(t)$ is the molecule position at time t. The MSD can be calculated from the particle position as a function of time, $x(t)$ (the position along the flow axis) and $y(t)$ (the position on the axis orthogonal to the flow). For our data, it is important to keep the diffusion coefficient along the flow axis ρ_x (calculated as $\rho_x(t) = [x(t) - x(0)]^2$ and along the orthogonal axis ρ_y separate, as there is both flow and diffusion in the x direction and only diffusion in the y direction, which furthermore behaves similarly to diffusion in a cage. The diffusion coefficient and flow rate can then be found by fitting the diffusion coefficient, D, and the velocity, V, from the measured curve using $\rho_x(t) = 2Dt + V^2t^2$, which is valid as long as the total measurement time is much smaller than $L^2/4D$, where L is the characteristic length of the space available for diffusion. This is valid along the channels in the x direction but not in the y direction owing to the presence of the channel walls[27].

smFRET data from SWIFT microscopy. Fluorescence resonance energy transfer efficiency (E_{FRET}) was calculated via

$$E_{FRET} = \frac{I_A^G}{I_D^G + I_A^G}$$

and stoichiometries (S) using

$$S = \frac{I_D^G + I_A^G}{I_D^G + I_A^G + I_A^O}$$

where I_x^y describes the background corrected intensity detected from the D or A dye and green (G) or orange (O) laser excitation. A background image was obtained by detecting particle positions from the sum of D and A channels over all images of the time series after alignment (based on fiducial beads). The image time series was then projected, and positions (spatial and temporal) where a particle was detected were not computed into the projection operation. This background image was then subtracted from the movie before data analysis.

21. Milles, S. & Lemke, E.A. *Biophys. J.* **101**, 1710–1719 (2011).
22. Lowary, P.T. & Widom, J. *J. Mol. Biol.* **276**, 19–42 (1998).
23. Lee, K.-M. & Narlikar, G. *Curr. Protoc. Mol. Biol.* **54**, 21.6 (2001).
24. Roy, I., Smith, O., Clouthier, C.M. & Keillor, J.W. *Protein Expr. Purif.* **87**, 41–46 (2013).
25. Gopich, I.V. & Szabo, A. *J. Phys. Chem. B* **111**, 12925–12932 (2007).
26. Vogelsang, J. *et al. Angew. Chem. Int. Ed. Engl.* **47**, 5465–5469 (2008).
27. Qian, H., Sheetz, M.P. & Elson, E.L. *Biophys. J.* **60**, 910–921 (1991).

ONLINE METHODS

Mouse liver study. The extended version of the mouse liver study has been recently published[10]. Here we analyzed only a subset of these data. Specifically, we focused on a group comparison of wild-type C57B6/J (B6) mice either fed *ad libitum* or fasted 12 h overnight, with three biological replicates in each condition. We used only the quantified transitions of the endogenous proteins as well as the available SILAC-labeled reference counterparts. The details of the selected subset of the data are as follows.

The full mouse liver study[10] was designed in three consecutive steps. First, light synthetic peptides were used to acquire MS2 spectra (AB Sciex QTrap 4000), build a reference spectral library and extract the most intense transitions and retention times for each peptide of interest. The peptides and the fragments in the library were targeted in the subsequent steps.

Second, several trial runs were used to determine which selected endogenous transitions had a reference SILAC counterpart. The trial runs used both unfractionated and fractionated liver homogenates with a heavy-labeled reference proteome (SILAC Hepa1-6 cell line). In this initial screening, 27 peptides (corresponding to 16 different proteins) did not have a SILAC reference counterpart.

Third, the actual experiment used a mixture of endogenous samples and SILAC-labeled reference samples. Prior to acquisition, heavy-labeled synthetic peptides were added for the 27 peptides missing the SILAC reference counterpart for quality control. The endogenous transitions of 24 of these targeted peptides were of a relatively good quality, and their quantification was unaffected by the presence of the synthetic reference peptides. The endogenous transitions of the remaining three targeted peptides (namely TVVSSEISGK, EASFVLHTISK and GISEVLART) were of a poor quality, regardless of the presence of the heavy channel. The raw files corresponding to the trial runs are publicly available in the PASSEL repository[21] (accession number PASS00322).

Here we analyzed the mouse liver data without taking the synthetic reference peptide into account. 12% (103/863) of reference transitions corresponding to 13% (16/122) reference proteins were missing in the SILAC-labeled samples, and 24% (207/863) of endogenous transitions and 19% (168/863) of reference transitions were missing in some runs for reasons other than by design.

Plots evaluating the accuracy of model-based imputation of missing reference transitions for every protein in the mouse liver study are shown in **Supplementary Data 1**. The list of identified and quantified transitions, which are used as input to label-sparse quantification of the mouse liver study, is available in **Supplementary Data 2**. The Skyline files of the screening (i.e., second) and experimental (i.e., third) steps of the experiment are available as **Supplementary Data 3** and **4**, respectively.

Pathway analysis (**Fig. 2**) was generated with Cytoscape[22]. All the physical or functional protein-protein interactions were obtained from the STRING database[23]. The proteins were further grouped on the basis of their known membership in KEGG pathways[24].

Gene Ontology enrichment analysis was performed using the DAVID functional annotation tool[25]. The Fisher exact test (equivalently, the hypergeometric test) was performed for the results of both label-sparse and full label-based approaches. The mouse proteome was selected as the universe, and the differentially abundant proteins (FDR cutoff of 0.01) were selected as the list of interest.

Additional technical details for label-sparse quantification. Full details of the label-sparse quantification approach are described in **Supplementary Note 1**. Full statistical details are available in **Supplementary Note 2**. Additional description of the experimental evaluation is available in **Supplementary Note 3**. An open-source software implementation is available at http://www.stat.purdue.edu/~ovitek/ and as **Supplementary Software**.

21. Farrah, T. *et al. Proteomics* **12**, 1170–1175 (2012).
22. Shannon, P. *et al. Genome Res.* **13**, 2498–2504 (2003).
23. Szklarczyk, D. *et al. Nucleic Acids Res.* **39**, D561–D568 (2011).
24. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. *Nucleic Acids Res.* **32**, D277–D280 (2004).
25. Huang, D.W., Sherman, B.T. & Lempicki, R.A. *Nat. Protoc.* **4**, 44–57 (2009).

ONLINE METHODS

Coverslips. Gold nanorods (25 × 45 nm, part #A12-25-650, Nanopartz) were deposited on the surface of poly(L-lysine) (Sigma)-coated 25-mm round coverslips of #1.5 thickness (Warner Instruments). SiO_2 (~20 nm thick) was sputtered over top of the nanorods using a Denton Explorer sputtering system (Denton Vacuum). These coverslips were cleaned and coated with poly(L-lysine) before the cells were plated.

Cell culture and transfection. PC12-GR5 cells were maintained at 37 °C and 5% CO_2 in T75 flasks (Sarsdtedt 83-1813-002) in dye-free DMEM (Life Technologies 21063-029) supplemented with 5% FBS and 5% horse serum. This line recently tested negative for mycoplasma contamination. Transfections were performed with Lipofectamine 2000 (Invitrogen) and 2 μg of plasmid after 1 d of growth on coverslips. The myristoylated psCFP2 expression constructed was made by PCR amplification of the psCFP2 gene (Evrogen), including the mammalian codon-optimized myristoylation signal peptide from the *SRC* gene: MGSSKSKPKDPSQRRNNN. The amplicon was subsequently cloned into the empty N1 mammalian expression vector (Clontech) using the BglII (5′) and NotI (3′) restriction sites. pEGFPC1-Epsin1 was obtained from Addgene (#22228). pmCherryC1-Epsin1 and pmEos3-Epsin1 were both made by replacing EGFP from pEGFPC1-Epsin1 with mCherry (Clontech #632524) or mEos3 (ref. 19) using Age1 and BsrG1. For live imaging, cells were transfected with 1 μg each of pEGFP-LCa (clathrin light chain A)[20] and pmCherryC1-Epsin1.

Unroofing. One day after transfection, cells were rinsed with imaging buffer (130 mM NaCl, 2.8 mM KCl, 5 mM $CaCl_2$, 1 mM $MgCl_2$, 10 mM HEPES, 10 mM glucose, pH 7.4), briefly incubated in one part stabilization buffer (70 mM KCl, 30 mM HEPES at pH 7.4, 5 mM $MgCl_2$ and 1 mM DTT) and three parts 0.01% (w/v) poly(L-lysine) for 10–15 s, and then incubated for 10–15 s three times in a fresh solution made of one part stabilization buffer and three parts water. The coverslips were then placed in stabilization buffer containing fixative (2% paraformaldehyde and 0.04% glutaraldehyde, Electron Microscopy Sciences) immediately before sonication. Sonication was done with a Branson Sonifier 450 with a 1/8" tapered microtip. The tip was positioned 5 mm above the coverslip, and a single 400-ms pulse at the lowest output setting resulted in approximately 1 cm^2 of unroofed cells on the coverslip. The cells were then placed into fresh stabilization buffer and fixed for 20 min.

Immunolabeling. Fixed unroofed cells were rinsed with PBS and placed into blocking buffer (PBS with 3% (w/v) bovine serum albumin, BSA) for 1 h. The unroofed cells were then incubated for 1 h in blocking buffer containing 2 μg/mL primary antibody (R-20, Santa Cruz Biotechnology, or X22, Thermo Scientific). The sample was rinsed with blocking buffer before secondary antibody labeling for 30 min (2 μg/mL Alexa Fluor 647 donkey anti-goat, Invitrogen A31571; Alexa Fluor donkey anti-mouse, Invitrogen A21447; or Atto 488 donkey anti-mouse) in blocking buffer. Atto 488 donkey anti-mouse was created with Atto 488 NHS ester (Sigma, 4-molar excess) and unlabeled donkey anti-mouse IgG (Abcam ab6707) and purified through a Superdex 75 10/300 GL size-exclusion column (GE Healthcare). Finally, the sample was washed with blocking buffer and PBS and was post-fixed for 20 min. For immunolabeling intact cells, 0.2% Triton X-100 was added to the blocking buffer, and cells were incubated for 2 min in permeabilization buffer (PBS with 3% (w/v) BSA, 0.5% Triton X-100) before blocking.

iPALM. After immunolabeling, cells were rinsed, placed in blinking buffer (50 mM Tris, 10 mM NaCl, 0.1 g/mL glucose, 0.8 mg/mL glucose oxidase, 40 μg/mL catalase, 71 mM 2-mercaptoethanol) and covered with an 18-mm #1.5 coverslip. The two coverslips were sealed with epoxy (Elmer's) and Vaseline (Unilever). iPALM was performed as previously described[6] with the *z*-axis measurement range extended to 750 nm (ref. 3). Typical iPALM data acquisition consisted of 40,000–80,000 frames acquired with iXon DU-897E EM-CCD cameras (Andor). Acquisition was performed in frame transfer mode, and the laser excitation was constantly active during acquisition. AF647 was imaged with a 637-nm MRL-III-640 laser (OptoEngine) at ~3,000 W/cm^2 with 20- to 30-ms exposure using LP02-647RU and FF01-720/SP filters. psCFP2 was imaged with 488-nm Cyan 488 laser excitation (Newport, Spectra Physics) at ~400 W/cm^2 with 50-ms exposure using LP02-488RS and FF01-520/35 filters. An additional quad notch filter NF01-405/488/561/635 was used for all imaging. Filters were from Semrock.

The iPALM calibration procedure and data analysis have been previously described[3,6,21]. Two-channel iPALM image registration was performed using Au nanoparticles. The same nanoparticles were observed under 647-nm illumination and 488-nm illumination, which allowed for registration of two channels[3]. Images of iPALM data were rendered as previously described[1,6].

After iPALM imaging of several areas (each containing 1–3 unroofed cells), a 3-mm circle was etched into the coverslip around the area with a diamond objective marker (#11505059, Leica Microsystems). The entire area was imaged with 10× differential interference contrast (DIC). The image was used as a map to locate cells in TEM. The coverslip sandwich was separated, and the sample was placed in HBSS (Hanks' balanced salt solution, Life Technologies) containing 2% glutaraldehyde at 4 °C overnight.

Electron microscopy. Samples were critical-point dried and were coated with platinum and carbon as previously described[22]. The coated samples were imaged with 10× phase contrast light microscopy to locate the regions of interest (ROIs). A pioloform and carbon–coated 50-mesh, 3-mm copper grid was plasma discharged and dipped in a 1:5 dilution of goat anti-mouse 10-nm immunogold conjugate (EM.GMTA10, BBinternational), rinsed and dried on filter paper. This resulted in sparsely scattered gold nanoparticles that were used as fiducials for tomogram alignment.

The platinum-carbon replica was lifted off of the coverslip by floating the sample on 5% hydrofluoric acid. The replicas were rinsed using successive dilutions with water, lifted out of the water and placed onto the grid using a Perfect Loop (Electron Microscopy Sciences). The replica was again imaged with 10× phase contrast light microscopy to find the placement of the ROIs with respect to the grid. In some cases, there was loss of a ROI because it was placed over a grid bar.

TEM was performed on a JEOL 1400 running SerialEM freeware[17] and equipped with a XR-111 CCD camera (Advanced

Microscopy Techniques) Montages of entire unroofed cells were produced at 15,000× with 10% overlap. Single-axis tilt series (−60° to 60°, 1° increments) were collected at 8,000×. The montages were stitched together, and the tilt series were reconstructed into tomograms using IMOD software[17,18].

Correlation. The 25 × 45–nm gold nanorods were identified in TEM micrographs. The procedure used for two-color PALM alignment[3] was also used to register two-color PALM images to the TEM images. For image correlations we used a POLYWARP1 image transformation defined as

$$X' = K_{x00} + K_{x01} \cdot X + K_{x10} \cdot Y + K_{x11} \cdot X \cdot Y$$
$$Y' = K_{y00} + K_{y01} \cdot X + K_{y10} \cdot Y + K_{y11} \cdot X \cdot Y$$

where X and Y are the PALM gold coordinates being transformed to the TEM gold coordinates, X' and Y'. K coefficients were calculated with a least-squared solution.

After *xy* PALM data were aligned to the 2D TEM data, the 10-nm gold nanospheres were used as markers to align the 2D TEM/iPALM correlation to the 3D tomogram (in the *xy* plane) using linear regression or POLYWARP1. *z* alignment was performed by manually aligning the membrane iPALM marker with the membrane in TEM tomograms across the entire tomogram.

Fluorescence probability density maps. 2D EM micrographs were analyzed in ImageJ[23], and elliptical regions of r_x and r_y (radii along the *x* and *y* axes) were drawn to best fit the shape of visible clathrin lattices. Clathrin structures were split into two categories: (i) flat or slightly curved domes and (ii) highly invaginated pits. Clathrin structures were omitted from this analysis only if they were at the edge of the cell, did not fit well to an ellipse, were within 0.5 μm of a gold nanorod or were right next to another clathrin structure. In Matlab (MathWorks), the coordinates of PALM localizations were mapped onto the coordinates of the ellipses and assigned a fractional position from $-2r_x$ to $2r_x$ in the *x* dimension and $-2r_y$ to $2r_y$ in the *y* dimension. The fractional coordinates from each ellipse were binned into a 40 × 40 2D histogram. The resulting image was normalized before the images from all regions of a specific category were averaged (**Figs. 2e**,**f** and **3b**,**c**). Radial scans were produced by averaging pixels on the basis of their distance away from the center of the image. The standard error of pixel sampling takes into account the total number of pixels included in the final data point (from each separate 2D histogram).

Geometric modeling. Fluorescence density maps were simulated in Matlab for two geometries: a hemisphere and a flat disk. We produced 2D projections (on a 1-nm grid) of these objects with radius r = 50 nm and a uniform shell of signal extending from their surfaces by 25 nm (accounting for the immunotether). The resulting image was convolved with Gaussians of variable widths. The width best simulating our data was σ = 16 nm. The final image was resampled to have 10-nm pixels (to simulate the fluorescence density maps produced by our data).

***z*-axis histograms.** The height of each structure was determined by selecting the local plasma membrane plane and the top of the clathrin structure in TEM tomograms. A 1D histogram of the *z* position of all iPALM localizations associated with each structure was mapped with respect to the TEM coordinates. Finally, histograms were aligned by their bottom TEM coordinate (the plasma membrane) and arranged in order of height (**Figs. 2i** and **3e**).

Two-color localization microscopy (for Supplementary Fig. 7). Unroofed cells were prepared as for correlative microscopy. Intact cells were rinsed with imaging buffer and fixed with 2% paraformaldehyde. Immunolabeling clathrin heavy chain with Atto 488 and epsin 1 with AF647 is described above. Fixed and immunolabeled cells were placed in blinking buffer and imaged with a Nikon N-STORM super-resolution microscope system (N-1245) equipped with a CFI SR Apochromat TIRF 100× oil objective, an 80-mW 488-nm laser, a 125-mW 647-nm laser and an Andor Ixon Ultra DU-897 camera. Two-color images were acquired in TIRF at full laser power with 20-ms frames. Final super-resolution images were localizations from 20,000–30,000 frames (each color).

AFM. Tapping-mode AFM images of clean coverslips were obtained in air with a multimode atomic force microscope driven by a NanoScope V controller and E scanner (Veeco/Bruker) using AC160TS cantilevers (Olympus) with a typical resonance frequency of 300 kHz and a nominal spring constant of 42 N/m. Images were analyzed using NanoScope software (version 7).

Live TIRF. Cells were grown on poly(L-lysine)-coated 25-mm round coverslips of #1.5 thickness (Warner Instruments) and transfected with pEGFP-LCa (clathrin light chain A)[20] and pmCherryC1-Epsin1. Cells were imaged by total-internal-reflection fluorescence (TIRF) in imaging buffer as described previously[22]. Each image shown is an average of five subsequent 100-ms frames each 500 ms apart.

19. Zhang, M. *et al. Nature Methods* **9**, 727–729 (2012).
20. Gaidarov, I., Santini, F., Warren, R.A. & Keen, J.H. *Nat. Cell Biol.* **1**, 1–7 (1999).
21. Kanchanawong, P. *et al. Nature* **468**, 580–584 (2010).
22. Sochacki, K.A. *et al. Nature Commun.* **3**, 1154 (2012).
23. Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. *Nat. Methods* **9**, 671–675 (2012).

doi:10.1038/nmeth.2816

ONLINE METHODS

Data sets. RA data (and statistical estimates of cell-type composition) were obtained directly from the authors of ref. 1 and are publicly available as indicated in their paper. They collected blood samples from 354 cases and 312 controls, which were assayed on an Illumina 450K DNA methylation chip. After filtering out failed probes and probes that were constitutively methylated or unmethylated (see "Data Processing") in all the samples, we retained 103,638 loci. Effects due to age, gender, smoking status and batch were removed by regressing these factors on methylation and then using the residuals for all further analysis (as in ref. 1). In ref. 1, the authors estimated the cell-type composition[9] for each sample by using carefully obtained reference methylation profiles (CD14 monocytes, CD19 B cells, CD4 T cells, CD56 NK cells and granulocytes) in conjunction with a statistical approach[9] and then added these as covariates to their association regression. We refer to this approach in the main text either as the bronze standard or as the reference-based approach; it was used to validate FaST-LMM-EWASher. Methylation and phenotype data from Lam *et al.*[4] were downloaded from the Gene Expression Omnibus (GEO) database (accession number GSE37008).

TCGA breast cancer DNA methylation data were downloaded (on 1 July 2013) from the TCGA data portal: https://tcga-data.nci.nih.gov/tcga/. The breast cancer methylation data set comprised 816 cases and 124 controls; we used case-control status as the phenotype. All methods (even the uncorrected) used batch and breast cancer subtypes luminal A, luminal B, basal and Her2 as covariates.

Data processing. Raw methylation data (from Illumina 450K or 27K arrays) were converted into β values. Following ref. 1, we deemed a site to be constitutively methylated if its average probe β value across all samples (cases and controls) was above 0.8 and to be constitutively unmethylated below 0.2. Because FaST-LMM-EWASher searches for loci correlated with the phenotype, we remove such constitutive probes from our association analysis for computational convenience. Linear regression was used to correct for known covariates such as batch (precise covariates are described with each data set description above), yielding residuals used in the main association analyses (for all methods).

Applications of the Houseman *et al.* methodology[9]. To obtain statistical estimates of the cell-type composition for the Lam *et al.*[4] blood-based methylation data, we used the R software from ref. 9 with the default 500 leukocyte differentially methylated regions (L-DMRs).

QQ plots and genomic control. Throughout our experiments we use quantile-quantile (QQ) plots of the $-\log_{10} P$ quantiles to assess inflation of the test statistic in our experiments, as is common in the GWAS community[16]. In these plots, quantiles of the theoretical null distribution are plotted against observed quantiles. Under the assumption that no methylation loci in the data are differentially expressed, the resulting plot should follow the diagonal and lie within the 95% confidence error bars (see **Supplementary Note 1** for computation of error bars). Because we expect some, but not too many methylation sites to be differentially expressed, we expect to see only small deviations from the 95% error bars, and we interpret greater deviations as inflation of the test statistic[16]. We also use the genomic control factor[17], λ, a summary statistic of the *P*-value distribution commonly used in GWAS to quantify how much the test statistics are inflated compared to the null distribution. The quantity λ is defined as the ratio of the median observed to median theoretical test statistic. A data set that corrects for confounders and contains only a small proportion of causal loci has λ around 1, whereas λ much greater than 1 indicates possible confounding and many false positives.

The FaST-LMM-EWASher algorithm. FaST-LMM-EWASher automatically finds the simplest combination of PCs with the LMM (the combination with the fewest number of PCs) that control for inflation of the test statistic. More technical details are available in **Supplementary Note 2**. The algorithm proceeds as follows:

1. Filter out loci that are constitutively high or low (see "Data Processing" above).
2. As in FaST-LMM-Select[7,8], run the uncorrected association analysis to rank all methylation loci by their significance. Then use the top K loci to construct the methylation similarity matrix, where K is automatically determined by maximizing out-of-sample cross-validation likelihood[8].
3. Using the similarity matrix determined from step 2 as the covariance component in the LMM, compute an association *P* value for each site. If the genomic control factor λ is still inflated (see note below), compute the PCs across all samples. Include the top PC as a covariate, and then repeat step 2: rerun the linear regression model to rerank all the loci by significance (now conditioned on the first PC), using the newly selected loci to construct the similarity matrix. This yields a FaST-LMM-EWASher model comprising the LMM and one PC, which is then used to compute association significance for each site. If λ is still inflated, repeat the process using the top two PCs as fixed covariates and iterate with an increasing number of PCs until the inflation is controlled.

The similarity matrix computed by FaST-LMM-EWASher correlates with (and serves as a proxy for) similarities between samples based on unobserved cell-type compositions (**Supplementary Fig. 9**). The matrix can also be used to visualize structures within samples (**Supplementary Fig. 10**). **Supplementary Figure 11** illustrates the results from FaST-LMM-EWASher's iterative model selection. On the colon cancer data set, FaST-LMM-EWASher with one or two PCs appears unable to correct for inflation of test statistics ($\lambda = 2.1$, and 1.1, respectively), whereas with three PCs, FaST-LMM-EWASher appears to sufficiently control inflation. For our data sets we set the inflation threshold to be $\lambda = 1$. Note that in GWAS with polygenic effects, it has been observed that λ can appear 'inflated' owing to signal in the data[18]; thus, setting the threshold of $\lambda = 1$ might be too conservative. In these cases, higher λ values can be tolerated. Similarly, in EWAS, if practitioners have prior belief that many loci might be truly associated with the phenotype beyond those tagging cell-type composition, then they can experiment by setting the FaST-LMM-EWASher λ threshold to be larger than 1 (i.e., the algorithm stops only after λ falls below this threshold). In studies where controlling for all

doi:10.1038/nmeth.2815

spurious associations is absolutely crucial, we suggest using the more conservative $\lambda = 1$ threshold.

Analogously to GWAS applications of the LMM[7,8], it is important to select, typically a subset of, methylation loci when constructing the similarity matrix. If we were to use all methylation loci to construct the similarity matrix, then loci in the genome that do not correlate with cell-type composition will introduce noise[7,8]. Therefore, we used feature selection based on the predictive log likelihood on samples held out by way of cross-validation, in conjunction with a univariate feature ranking from linear regression, as in ref. 8. When we omitted this step, using all loci to compute similarity on the RA data, inflation of the test statistic was not sufficiently corrected for (see **Supplementary Fig. 12a**). Similarly, it is important to reselect loci after adding additional PC covariates; not doing so yielded inflated test statistics (**Supplementary Fig. 12c,d**). Combining PCs with LMMs has been suggested in the context of GWAS[19], although, to our knowledge, it has not fully been explored. The idea of performing feature selection for features in the LMM similarity matrix, iteratively, as we condition on increasingly more PCs covariates, is a novel contribution of this paper, considerably improving the performance of FaST-LMM-EWASher.

Parameter estimation and statistical testing. Parameters in the linear mixed model were estimated by way of maximum likelihood (ML). Although restricted maximum likelihood (REML) can provide benefits over ML, this is true only for small sample sizes or when many covariates are used. We did not find differences between REML and ML in experiments for this paper. Association *P* values for all models used (linear regression with or without covariates, linear mixed model with or without covariates) were obtained by using a 1 degree-of-freedom likelihood-ratio test, where the null model did not include the methylation locus being tested, and the alternative model did. All Bonferroni-significant tests were performed at the nominal significance level $\alpha = 0.05$.

Simulations. To simulate null-only data in such a way as to most closely mimic the actual RA data, we obtained the previously inferred cell-type composition for each sample in RA (inferred using software provided with ref. 9) as well as the reference DNA methylation profiles for each of the five blood cell types[1]. We used the same number of probes and the same number of cases and controls as in the original RA data. To simulate a sample (a methylation profile for one individual), we took the weighted average of the five reference cell types, with weights given by the previously inferred cell-type composition on the real data. We then added independent and identically distributed Gaussian noise to each locus. The performance of FaST-LMM-EWASher and the reference-based cell-type composition method was robust to the amount of noise added (we tried noise s.d. in the range 0.05–0.3). Figures in **Supplementary Figure 5** were generated with noise s.d. of 0.1. We simulated locus-specific signal by selecting a cell type (or a set of cell types) as being differentially methylated between case and control and then creating a case (or control) reference methylation profile for that cell type by making the causal loci systematically higher (or lower) compared to the normal reference profile. The samples were then simulated as before, only now using a weighted sum of either the case or control reference profiles, as appropriate.

Enrichment analysis. For GO enrichment analysis, we first assigned loci to genes using the University of Santa Cruz (UCSC) Genome Browser annotations (build hg19). If a locus was in the promoter or gene body, we assigned it to that gene. Intergenic loci were not assigned to genes. For the GO analysis, we identified those genes associated with the top 100 markers from each method (FaST-LMM-EWASher and the uncorrected analysis) and then performed gene-set enrichment analysis on each of these two sets of genes using DAVID[20,21], reporting the most significantly enriched categories based on *P* values reported by DAVID.

Software. Python and R software associated with our method (available from http://www.microsoft.com/science) is built upon and extends FaST-LMM[6–8], which performs the linear mixed model analysis (including feature selection). Our suite of tools includes functionalities to cluster and visualize the similarity matrix as a heat map, in addition to performing the FaST-LMM-EWASher association analysis.

FaST-LMM-EWASher is computationally efficient. Each of the data sets analyzed (RA, Lam *et al.*, TCGA breast and colon cancer) were analyzed by FaST-LMM-EWASher on a single laptop (Lenovo X1 Carbon with 8 GB of RAM) in 1–5 min. Furthermore, the FaST-LMM linear mixed model backbone of FaST-LMM-EWASher has been successfully scaled to large GWAS data sets[6–8], and the FaST-LMM-EWASher running time is just a constant (equal to the number of PCs scanned, typically 1–3) times the LMM running time. Therefore, as the size of EWAS data sets approaches that of GWAS, FaST-LMM-EWASher will be a fast and memory-efficient tool performing genome-wide analyses.

16. Balding, D.J. *Nat. Rev. Genet.* **7**, 781–791 (2006).
17. Devlin, B. & Roeder, K. *Biometrics* **55**, 997–1004 (1999).
18. Yang, J. *et al. Eur. J. Hum. Genet.* **19**, 807–812 (2011).
19. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. *Nat. Rev. Genet.* **11**, 459–463 (2010).
20. Huang, D.W., Sherman, B.T. & Lempicki, R.A. *Nucleic Acids Res.* **37**, 1–13 (2009).
21. Huang, D.W., Sherman, B.T. & Lempicki, R.A. *Nat. Protoc.* **4**, 44–57 (2009).

ONLINE METHODS

Materials. Unmodified DNA oligonucleotides were purchased from Integrated DNA Technologies. Fluorescently modified DNA oligonucleotides were purchased from Biosynthesis. Biotinylated monoclonal antibodies against β-tubulin (9F3; catalog number: 6181) and COX IV (3E11; catalog number: 6014) were purchased from Cell Signaling. Anti-PMP70 (catalog number: ab28499) was purchased from Abcam. Anti-TGN46 (catalog number: NBP1-49643B) was purchased from VWR. Streptavidin was purchased from Invitrogen (catalog number: S-888). Bovine serum albumin (BSA) and BSA-biotin obtained from Sigma-Aldrich (catalog number: A8549). Glass slides and coverslips were purchased from VWR. Lab-Tek II chambered coverglass was purchased from Thermo Fisher Scientific. M13mp18 scaffold was obtained from New England BioLabs. p8064 scaffold for microtubule-like DNA origami structures was prepared as described before[19]. Freeze 'N Squeeze columns were ordered from Bio-Rad. TetraSpeck Beads were purchased from Life Technologies. Paraformaldehyde, glutaraldehyde and TEM grids (FORMVAR 400 mesh copper grids) were obtained from Electron Microscopy Sciences.

Three buffers were used for sample preparation and imaging: buffer A (10 mM Tris-HCl, 100 mM NaCl, 0.05% Tween 20, pH 7.5), buffer B (5 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM EDTA, 0.05% Tween 20, pH 8) and buffer C (1× PBS, 500 mM NaCl, pH 8).

Optical setup. Fluorescence imaging was carried out on an inverted Nikon Eclipse Ti microscope (Nikon Instruments) with the Perfect Focus System, applying an objective-type TIRF configuration using a Nikon TIRF illuminator with an oil-immersion objective (CFI Apo TIRF 100×, numerical aperture (NA) 1.49, oil). For 2D imaging an additional 1.5× magnification was used to obtain a final magnification of ~150-fold, corresponding to a pixel size of 107 nm. Three lasers were used for excitation: 488 nm (200 mW nominal, Coherent Sapphire), 561 nm (200 mW nominal, Coherent Sapphire) and 647 nm (300 mW nominal, MBP Communications). The laser beam was passed through cleanup filters (ZT488/10, ZET561/10 and ZET640/20, Chroma Technology) and coupled into the microscope objective using a multiband beam splitter (ZT488rdc/ZT561rdc/ZT640rdc, Chroma Technology). Fluorescence light was spectrally filtered with emission filters (ET525/50m, ET600/50m and ET700/75m, Chroma Technology) and imaged on an electron-multiplying charge-coupled device (EMCCD) camera (iXon X3 DU-897, Andor Technologies).

DNA origami self-assembly. The microtubule-like DNA origami structures were formed in a one-pot reaction with a 40-µl total volume containing 10 nM scaffold strand (p8064), 500 nM folding staples and biotin handles, 750 nM biotin anti-handles and 1.1 µM DNA-PAINT docking strands in folding buffer (1× TAE buffer with 20 mM $MgCl_2$). The solution was annealed using a thermal ramp[13] cooling from 80 °C to 14 °C over the course of 15 h. After self-assembly, monomeric structures were purified by agarose gel electrophoresis (1.5% agarose, 0.5× TBE, 10 mM $MgCl_2$, 1× SybrSafe) at 4.5 V/cm for 1.5 h (see **Supplementary Fig. 2**). Gel bands were cut, crushed and filled into a Freeze 'N Squeeze column and spun for 5 min at 1,000*g* at 4 °C. Polymerization was carried out at 30 °C for 48 h with a fivefold excess of polymerization staples in folding buffer. Polymerized structures were used for imaging without further purification.

DNA origami drift markers were self-assembled in a one-pot reaction (40-µl total volume, 20 nM M13mp18 scaffold, 100 nM biotinylated staples, 530 nM staples with DNA-PAINT docking sites, 1× TAE with 12.5 mM $MgCl_2$). Self-assembled structures were purified as described before.

DNA origami structures for the four-color *in vitro* Exchange-PAINT demonstration were self-assembled in a one-pot reaction (40-µl total volume, 30 nM M13mp18 scaffold, 470 nM biotinylated staples, 400 nM staples with docking sites for number imaging, 370 nM core structure staples, 1× TAE with 12.5 mM $MgCl_2$). Self-assembled structures were purified as described before.

DNA origami structures for the ten-color *in vitro* Exchange-PAINT demonstration were self-assembled in a one-pot reaction (40-µl total volume, 30 nM M13mp18 scaffold, 36 nM biotinylated staples, 750 nM staples with docking sites for number imaging, 300 nM core structure staples, 1× TAE with 12.5 mM $MgCl_2$). Structures were not purified. Excessive staples were washed out of the sample after immobilization of the structure on the surface.

DNA strand sequences for the microtubule-like DNA origami structures can be found in **Supplementary Table 1**. DNA strand sequences for DNA origami drift markers can be found in **Supplementary Table 2**. DNA strand sequences for DNA origami structures for the ten-color *in vitro* Exchange-PAINT demonstration can be found in **Supplementary Tables 3** and **4** for odd and even digits, respectively. DNA strand sequences for DNA origami structures for *in vitro* Exchange-PAINT demonstration (digits 0–3) can be found in **Supplementary Table 5**. The scaffold sequence for p8064 and M13mp18 can be found in **Supplementary Tables 6** and **7**, respectively.

DNA-PAINT imager and docking sequences as well as sequences for surface attachment via biotin are listed in **Supplementary Table 8**.

Antibody-DNA conjugates. Antibody-DNA conjugates used to specifically label proteins of interest with DNA-PAINT docking sites were preassembled in two steps. First, 3.2 µl of 1 mg/ml streptavidin (dissolved in buffer A) was reacted with 0.5 µl biotinylated DNA-PAINT docking strands at 100 µM and an additional 5.3 µl of buffer A for 30 min at room temperature (RT) while gently shaking. The solution was then incubated in a second step with 1 µl of monoclonal biotinylated antibodies at 1 mg/ml against the protein of interest for 30 min at RT. Filter columns (Amicon 100 kDa, Millipore) were used to purify the preassembled conjugates from unreacted streptavidin-oligo conjugates.

Cell immunostaining. HeLa and DLD1 cells were cultured with Eagle's minimum essential medium fortified with 10% FBS with penicillin and streptomycin and were incubated at 37 °C with 5% CO_2. At approximately 30% confluence, cells were seeded into Lab-Tek II chambered coverglass 24 h before fixation. Microtubules, mitochondria, Golgi complexes and peroxisomes were immunostained using the following procedure: washing in PBS; fixation in a mixture of 3% paraformaldehyde and 0.1% glutaraldehyde in PBS for 10 min; 3× washing with PBS; reduction with ~1 mg/ml $NaBH_4$ for 7 min; 3× washing with PBS; permeabilization with 0.25% (v/v) Triton X-100 in PBS for 10 min; 3× washing with PBS; blocking with 3% (w/v) BSA for 30 min

and staining overnight with the preassembled antibody-DNA conjugates against β-tubulin, COX IV, PMP70 or TGN46 (conjugates were diluted to 10 μg/ml in 5% BSA); 3× washing with PBS; postfixation in a mixture of 3% paraformaldehyde and 0.1% glutaraldehyde in PBS for 10 min; and 3× washing with PBS.

Super-resolution DNA-PAINT imaging of microtubule-like DNA origami structures. For sample preparation, a piece of coverslip (no. 1.5, 18 × 18 mm^2, ~0.17 mm thick) and a glass slide (3 × 1 $inch^2$, 1 mm thick) were sandwiched together by two strips of double-sided tape to form a flow chamber with inner volume of ~20 μl. First, 20 μl of biotin-labeled bovine albumin (1 mg/ml, dissolved in buffer A) was flown into the chamber and incubated for 2 min. The chamber was then washed using 40 μl of buffer A. 20 μl of streptavidin (0.5 mg/ml, dissolved in buffer A) was then flown through the chamber and allowed to bind for 2 min. After washing with 40 μl of buffer A and subsequently with 40 μl of buffer B, 20 μl of biotin-labeled microtubule-like DNA structures (~300 pM monomer concentration) and DNA origami drift markers (~100 pM) in buffer B were finally flown into the chamber and incubated for 5 min. The chamber was washed using 40 μl of buffer B.

The final imaging buffer solution contained 1.5 nM Cy3b-labeled imager strands in buffer B. The chamber was sealed with epoxy before subsequent imaging. The CCD readout bandwidth was set to 1 MHz at 16 bit and 5.1 pre-amp gain. No electron-multiplying (EM) gain was used. Imaging was performed using TIR illumination with an excitation intensity of 294 W/cm^2 at 561 nm.

Super-resolution Exchange-PAINT imaging of DNA nanostructures. For fluid exchange, a custom flow chamber was constructed as shown in **Supplementary Figure 6a**. A detailed preparation protocol can be found in the **Supplementary Protocol**. Prior to the functionalizing of the imaging chamber with BSA-biotin, it was rinsed with 1 M KOH for cleaning. Binding of the origami structures to the surface of the flow chamber was performed as described before. Each image acquisition step was followed with a brief ~1–2 min washing step consisting of at least three washes using 200 μl of buffer B for each. Then the next imager strand solution was introduced. The surface was monitored throughout the washing procedure to ensure complete exchange of imager solutions. Acquisition and washing steps were repeated until all ten targets were imaged. The CCD readout bandwidth was set to 3 MHz at 14 bit and 5.1 pre-amp gain. No EM gain was used. Imaging was performed using TIR illumination with an excitation intensity of 166 W/cm^2 at 561 nm (ten-color Exchange-PAINT with 3 nM Cy3b-labeled imager strands in buffer B; **Fig. 3c**,**d**) and 600 W/cm^2 at 647 nm (four-color Exchange-PAINT with 3 nM Atto 655–labeled imager strands in buffer B; **Fig. 3e**).

Super-resolution DNA-PAINT imaging of cells. All data were acquired with an EMCCD readout bandwidth of 5 MHz at 14 bit, 5.1 pre-amp gain and 255 EM gain. Imaging was performed using HILO illumination[11]. The laser power densities were 283 W/cm^2 at 647 nm in **Figure 2a** and 142 W/cm^2 at 647 nm and 19 W/cm^2 at 561 nm in **Figure 2d**.

Imaging conditions were as follows. For **Figure 2a** we used 700 pM Atto 655–labeled imager strands in buffer C. For **Figure 2d** we used 600 pM Cy3b-labeled imager strands and 1.5 nM Atto 655–labeled imager strands in buffer C.

Super-resolution Exchange-PAINT imaging of cells. A Lab-Tek II chamber was adapted for fluid exchange as shown in **Supplementary Figure 6b**. 2D images (**Fig. 4a**, **i–iv**) were acquired with an EMCCD readout bandwidth of 5 MHz at 14 bit, 5.1 pre-amp gain and 255 EM gain. 3D images (**Fig. 4b–d**) were acquired with a CCD readout bandwidth of 3 MHz at 154 bit, 5.1 pre-amp gain and no EM gain. Imaging was performed using HILO illumination in both cases. Sequential imaging was done as described for the 2D origami nanostructures, but the washing steps were performed using buffer C. The laser power densities at 647 nm were 257 W/cm^2 in **Figure 4a**, **i** and 385 W/cm^2 in **Figure 4a**, **ii–iv**. The laser power densities at 561 nm were 31 W/cm^2 in **Figure 4b–d**.

Imaging conditions were as follows. For **Figure 4a**, **i** we used 700 pM Atto 655–labeled imager strands in buffer C. For **Figure 4a**, **ii–iv** we used 2 nM Atto 655–labeled imager strands in buffer C. For **Figure 4b** we used 800 pM Cy3b-labeled imager strands in buffer C. For **Figure 4c**,**d** we used 2 nM Cy3b-labeled imager strands in buffer C.

3D DNA-PAINT imaging. 3D images were acquired with a cylindrical lens in the detection path (Nikon). The N-STORM analysis package for NIS Elements (Nikon) was used for data processing. Imaging was performed without additional magnification in the detection path, yielding 160-nm pixel size. 3D calibration was carried out according to the manufacturer's instructions.

Imager strand concentration determination. Optimal imager concentrations were determined empirically according to the labeling density. Generally, a high enough fluorescence off-on ratio has to be ensured in order to guarantee binding of only a single imager strand per diffraction-limited area. Additionally, a sufficient imager strand concentration (and thus sufficiently low fluorescence off-time) is necessary to ensure sufficient binding events and thereby robust detection of every docking strand during image acquisition.

Super-resolution data processing. Super-resolution DNA-PAINT images were reconstructed using spot-finding and 2D-Gaussian fitting algorithms programmed in LabVIEW[10] (**Supplementary Software**). A simplified version of this software is available for download at http://www.dna-paint.net/ or http://molecular-systems.net/software/.

Normalized cross-correlation analysis. Normalized cross-correlation coefficients were obtained by first normalizing the respective reconstructed grayscale super-resolution images and subsequently performing a cross-correlation analysis in Matlab R2013b (MathWorks).

Drift correction and channel alignment. DNA origami structures (**Supplementary Fig. 1**) were used for drift correction and as alignment markers in *in vitro* DNA-PAINT and Exchange-PAINT imaging. Drift correction was performed by tracking the position of each origami drift marker throughout the duration of each image acquisition. The trajectories of all detected drift

doi:10.1038/nmeth.2835

markers were then averaged and used to globally correct the drift in the final super-resolution reconstruction. For channel alignment between different imaging cycles in Exchange-PAINT, these structures were used as alignment points by matching their positions in each Exchange-PAINT image.

For cellular imaging, 100-nm-diameter gold nanoparticles (Sigma-Aldrich;10 nM in buffer C, added before imaging) were used as drift and alignment markers. The gold nanoparticles adsorb nonspecifically to the glass bottom of the imaging chambers. Drift correction and alignment was performed in a fashion similar to that for the origami drift markers. Again, the apparent movement of all gold nanoparticles in a field of view was tracked throughout the image acquisition. The obtained trajectories were then averaged and used for global drift correction of the final super-resolution image. For the dual-color image of mitochondria and microtubules in **Figure 2d–f**, the gold particles were visible in both color channels. The same gold nanoparticles were also used for drift correction and realignment of the different imaging rounds in the *in situ* Exchange-PAINT experiments (**Fig. 4**).

Transmission electron microscopy (TEM) imaging. For imaging, 3.5 µl of undiluted microtubule-like DNA structures were adsorbed for 2 min onto glow-discharged, carbon-coated TEM grids. The grids were then stained for 10 s using a 2% aqueous ultrafiltrated (0.2-µm filter) uranyl formate solution containing 25 mM NaOH. Imaging was performed using a JEOL JEM-1400 operated at 80 kV.

Atomic force microscopy imaging. Imaging was performed using tapping mode on a Multimode VIII atomic force microscope (AFM) with an E-scanner (Bruker). Imaging was performed in TAE/Mg^{2+} buffer solution with DNP-S oxide-sharpened silicon nitride cantilevers and SNL sharp nitride levers (Bruker Probes) using resonance frequencies between 7 and 9 kHz of the narrow 100-µm, 0.38–N/m–force constant cantilever. After self-assembly of the origami structure, ~20 µl of TAE/Mg^{2+} buffer solution was deposited onto a freshly cleaved mica surface (Ted Pella) glued to a metal puck (Ted Pella). After 30 s the mica surface was dried using a gentle stream of N_2, and 5 µl of the origami solution was deposited onto the mica surface. After another 30 s, 30 µl of additional buffer solution was added to the sample. Imaging parameters were optimized for best image quality while the highest possible set point was maintained to minimize damage to the samples. Images were postprocessed by subtracting a first-order polynomial from each scan line. Drive amplitudes were approximately 0.11 V, integral gains were ~2 and proportional gains were ~4.

doi:10.1038/nmeth.2835

ONLINE METHODS

HeLa cell culture. Human epithelial carcinoma cells of the line HeLa (ATCC, S3 subclone) were cultured in SILAC DMEM where applicable (PAA Laboratories, E15-086), supplemented with 10% dialyzed FBS (PAA Laboratories, A15-107), 20 mM glutamine (PAA Laboratories, M11-006), 1% penicillin-streptomycin (PAA Laboratories, P11-010), 42 mg/l L-arginine (Sigma-Aldrich, A6969) and 62 mg/l L-lysine (Sigma-Aldrich, L8662). Cells were tested for mycoplasma contamination. For preparation of heavy isotope–labeled peptides, medium contained 42 mg/l [$^{13}C_6{}^{15}N_4$]arginine (Arg^{10}, Cambridge Isotope Laboratories, CNLM-539) and 61 mg/l [$^{13}C_6{}^{15}N_2$]lysine (Lys^8, Cambridge Isotope Laboratories, CNLM-291) instead of the natural amino acids. Cells were cultured for six passages until they were fully labeled. The cells were collected by centrifugation at 200*g* for 10 min, washed once with cold PBS and resuspended in cold PBS. Cell viability and number counts were performed according to the manufacturer using a Countess Automated Cell Counter (Life technologies, C10227).

Yeast cell culture. Budding yeast (*S. cerevisiae*) strains BY4741, YBR115C (*lys2* deletion strain) and fission yeast (*S. pombe*) strain SP286 were acquired from EUROSCARF, Thermo Scientific or Bioneer, respectively. The wild-type strain BY4741 was grown in YPD medium (20 g/l Bacto peptone (BD, 211677), 10 g/l yeast extract (Fisher Scientific, BP1422-2)) supplemented with 2% w/v glucose (Sigma-Aldrich, G7021). SILAC labeling of YBR115C was achieved by growing the cells for at least eight generations in SC medium supplemented with L-$^{13}C_6{}^{15}N_2$-lysine (Cambridge Isotope Laboratories, CNLM-291) and 2% w/v glucose. Fission yeast was grown in YES medium (5 g/l Bacto yeast extract supplemented with 3% w/v glucose and 250 mg/l of each adenine (Sigma-Aldrich, A2786), L-histidine (Sigma-Aldrich, H6034), L-leucine (Sigma-Aldrich, L8912), uracil (Sigma-Aldrich, U1128) and L-lysine (Sigma-Aldrich, L862)). Cells were grown at 30 °C to an OD_{600} of 0.6, harvested by centrifugation at 500*g* for 5 min at 4 °C, washed once with water and stored at −80 °C.

Tryptophan fluorescence emission assay for protein quantification. Protein concentrations were determined by tryptophan fluorescence emission at 350 nm using an excitation wavelength of 295 nm. Briefly, 1 µl of sample was solubilized in 200 µl of 8 M urea, and tryptophan at a concentration of 0.1 µg/µl was used to build a standard calibration curve (0.25–1.5 µl). Protein concentration in samples was estimated considering the emission of 0.1 µg/µl tryptophan equivalent to the emission of 7 µg/µl of human protein extract, assuming that tryptophan accounts for 1.3% on the human protein amino acid composition, on average.

in-StageTip lysis, reduction and alkylation. Quantities of up to 20 µg protein material were loaded directly onto the enclosed StageTips (Eppendorf epT.I.P.S., 0030073266); larger quantities were lysed and digested in a separate vial before loading a StageTip. To avoid clogging, typically 14-gauge StageTip plugs were used. Unless otherwise stated, approximately 10 µg or 20 µg protein starting material was used for single-shot or fractionation sample preparations, respectively (**Fig. 2**, **Supplementary Video 1** and **Supplementary Table 3**). Cells were lysed in lysis buffer (**Supplementary Table 3**) at a ratio of 1–5 µg protein per 1 µl lysis buffer (*S. cerevisiae*, *S. pombe* and HeLa cells contain approximately 3 pg/cell, 9 pg/cell and 200 pg/cell of protein, respectively). To simplify calculation, yeast cells corresponding to 1 ml culture at OD_{600} = 1 should be lysed in 60 µl lysis buffer, and 10^6 HeLa cells should be lysed in 300 µl lysis buffer. The lysates were boiled for 5 min and then sonicated to denature proteins, shear DNA and enhance cell disruption using a water-bath sonicator for enclosed StageTips (Bioruptor, model UCD-200, Diagenode) for 15 min at level 5, or a Sonifier for large volumes (>500 µl) (model 250, Branson Ultrasonics) for 1 min at duty cycle 20% and output control 3. If bead-milling is desired, an adaptor for a bead-milling system (MP Biomedicals, FastPrep-24) can be constructed by drilling a centered 2 mm diameter hole at the bottom of a 2 ml screw-cap micro tube (Sarstedt AG & Co., 72.694.006). The enclosed StageTip filled with ~100 µl beads (Lysing Matrix Y, MP Biomedicals) can be placed inside the bead-milling adaptor.

Lysates were diluted for digestion using a dilution buffer (**Supplementary Table 3**). The dilution buffer should contain respective amounts of proteolytic enzyme to ensure a ratio of 1:50 (micrograms of enzyme to micrograms of protein). Digestion was performed at 37 °C overnight. Peptides were acidified for C_{18}, SDB-XC, SDB-RPS and SCX materials and basified for SAX material (**Supplementary Table 3**). The StageTips were centrifuged using an in-house-made StageTip centrifuge (identical specifications to the Sonation StageTip centrifuge) for up to 2,000*g*; for higher centrifugation speed, Eppendorf tube adaptors (STH01, Sonation) were used. The StageTip was washed 1–3 times using 100 µl washing buffer depending on the number of plugs (**Supplementary Table 3**). Elutions were performed using 60 µl elution buffer depending on the StageTip material and whether fractionation was intended (**Supplementary Table 3**). All eluted materials were collected in autosampler vials and dried using a SpeedVac centrifuge at room temperature (Eppendorf, Concentrator plus, 5305 000.304). If remnants were visible after drying, the pellet was resuspended in double-distilled water followed by a second drying step. Only SAX elutions needed additional C_{18} desalting. Peptides were resuspended in buffer A* (2% acetonitrile (ACN) and 0.1% trifluoroacetic acid (TFA)) and were briefly sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510).

96-well processing device. A 96-well StageTip holder (length × width × height: 127 mm × 85 mm × 32 mm) was designed by drilling conical holes with the measures of the pipette tips (4 mm diameter) in a spacing corresponding to a 96-well PCR plate (9 mm). The material of the block was polyoxymethylene (POM). A second holder for PCR tubes was designed using a plate of the same material and equal area but 5 mm height. In the second holder, holes were drilled with equal spacing to hold PCR tubes (5 mm diameter). Spacers of 11 mm height above and 6 mm below the PCR holder were placed in the corners to of the two plates to maintain the distance between the StageTip holder and the PCR tube holder and to guarantee a correct alignment of the StageTips and the PCR tubes.

Phase transfer surfactant-aided in-solution sample preparation. In-solution sample preparation was performed as previously described[40] with some adjustments for a more comprehensive comparison of the method. In brief, 1,000, 10,000 and 100,000

HeLa cells were resuspended in 5 µl 1% (w/v) sodium deoxycholate, 10 mM TCEP, 40 mM 2-chloroacetamide (CAA), 100 mM Tris, pH 8.5, and subsequently lysed by 5 min boiling at 95 °C and sonication (Bioruptor, model UCD-200, Diagenode) for 15 min at level 5. Cell debris were pelleted by centrifugation at 13,200 r.p.m. for 5 min and the clarified lysate was transferred into a new vial. The lysate was diluted 1:10 for LysC-trypsin digestion (0.4 µg of each enzyme in double distilled water), and the digestion was performed overnight at 37 °C. The digest was acidified with 50 µl 2% TFA and sodium deoxycholate was extracted using 50 µl ethyl acetate and vigorous shaking. The organic phase was removed after centrifugation at 13,200 r.p.m. for 5 min. Finally, the peptides were desalted on C_{18} StageTips (**Supplementary Table 3**). The LC-MS set up used for in-solution experiments was the same as described below.

Liquid chromatography and mass spectrometry. Approximately 1 µg or 2 µg of peptides were loaded for 2 h or 4 h gradients, respectively. Peptides were separated on a 50-cm 75-µm inner diameter column packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH). Reverse-phase chromatography was performed with an EASY-nLC 1000 ultra-high pressure system (Thermo Fisher Scientific), which was coupled to the Q Exactive mass spectrometer (Thermo Fisher Scientific) via a nanoelectrospray source (Thermo Fisher Scientific). Peptides were loaded in buffer A (0.1% (v/v) formic acid) and eluted with a non-linear 120-min or 240-min gradient of 5–60% buffer B (0.1% (v/v) formic acid, 80% (v/v) acetonitrile) at a flow rate of 250 nl/min. After each gradient, the column was washed with 95% buffer B for 3 min and reequilibrated with buffer A for 3 min. Column temperature was kept at 50 °C by an in-house–designed oven with a Peltier element and operational parameters were monitored in real time by the SprayQc software[41]. MS data were acquired with an automatic switch between a full scan and up to five or ten data-dependent MS/MS scans (topN method). Target value for the full scan MS spectra was 3×10^6 charges in the 300–1,700 *m/z* range with a maximum injection time of 20 ms and a resolution of 70,000 at *m/z* 400. Isolation of precursors was performed with a 1.6 *m/z* window and a fixed first mass of 100.0 *m/z*. Precursors were fragmented by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 25 eV. MS/MS scans were acquired at a resolution of 17,500 at *m/z* 400 with an ion target value of 1×10^6 and a maximum injection time of 60 ms. Repeat sequencing of peptides was minimized by excluding the selected peptide candidates for 45 s.

Data analysis. MS raw files were analyzed by MaxQuant software (version 1.3.10.12) and peak lists were searched either against the human Uniprot FASTA database version of 25 February 2012 (81213 entries), against the *S. cerevisiae* Uniprot FASTA database version of 25 February 2012 (6,649 entries) or the *Saccharomyces* genome database–based *S. cerevisiae* FASTA database orf_trans.20100105 (5,904 entries) or against the *S. pombe* Uniprot FASTA database version of 2 April 2013 (5,096 entries) or *S. pombe* FASTA database version of 2 April 2013 (5,031 entries) and a common contaminants database (247 entries) by Andromeda search engine[42] with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. False discovery rate was set to 0.01 for proteins and peptides (minimum length of 7 amino acids) and was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine, and a maximum of two missed cleavages were allowed in the database search. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 p.p.m. and an allowed fragment mass deviation 20 p.p.m. Quantification of SILAC pairs was carried out by MaxQuant with standard settings and without the requantification option.

Bioinformatics analysis. Data analysis was performed with Perseus software in the MaxQuant computational platform and by R statistical computing environment. All enrichment analysis and analysis of variance tests were performed with Benjamini-Hochberg correction at a false discovery rate of 0.02.

Absolute quantification of protein abundances (copy numbers) were computed using peptide label-free quantification values, sequence length and molecular weight as described before[22] based on a total protein per cell value of 3 pg, 10 pg or 200 pg for *S. cerevisiae*, *S. pombe* or HeLa cells, respectively.

To assign protein orthologs between *S. cerevisiae*, *S. pombe* and HeLa cells, the Uniprot identifier was annotated with its corresponding eggNOG identifier[28]. In case of the same eggNOG identifier for multiple Uniprot identifiers, the median copy number for the corresponding protein groups was calculated. The resulting data set contained information about the UniProt identifier of the identified protein groups, the protein and gene name, the copy number as well as the eggNOG identifier, indicating orthologs between *S. cerevisiae*, *S. pombe* and HeLa cells (**Supplementary Table 2**).

To analyze GO-term differences between orthologs, we used the 2D Annotation Enrichment technique[18] that employs a two-dimensional generalization of the nonparametric two-sample test and uses the Benjamini-Hochberg false discovery rate to correct for multiple-hypotheses testing.

40. Masuda, T., Tomita, M. & Ishihama, Y. Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J. Proteome Res.* **7**, 731–740 (2008).
41. Scheltema, R.A. & Mann, M. SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* (11 May 2012).
42. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

doi:10.1038/nmeth.2834

ONLINE METHODS

Construction of transgenic animals. Plasmids were constructed by standard DNA cloning and PCR methods. All PCR reactions were performed using PrimeStar HS DNA polymerase (Takara). Following amplification, all sequences were verified by DNA sequencing.

UAS-ChR2(H134R)::EYFP-2A-ChR2(H134R)::EYFP. A DNA fragment containing the ChR2(H134R) coding sequence, kindly provided by K. Deisseroth, and an intervening F2A sequence[12,41] were amplified by PCR using primers (5F-EcoRI-chr2, 3R-2a-YFP, 5F-2a-Chr2, 3R-Xba-YFP, 5F-2a and 3R-2a) and subcloned into pUAST vector in a tandem manner using restriction enzymes (see **Supplementary Table 4** for primer sequences). Several transgenic flies were created with different insertion sites. We picked the line that exhibited the strongest induction of PER when crossed to Gr5a-*GAL4*.

UAS-C1V1(T/T). A DNA fragment containing the coding sequence of C1V1(E122T/E162T)-TS-EYFP kindly provided by K. Deisseroth was amplified by PCR using primers (C1V1-f and C1V1-EGFP-r; see **Supplementary Table 4**). This PCR product was subcloned into the vector pJFRC2 (ref. 30) using SLIC cloning[42]. This vector was injected and integrated into attP40 and VK5 sites[30].

UAS-ReaChR, LexAop-ReaChR, UAS-FRT-*mCherry*-FRT-*ReaChR and LexAop*-FRT-*mCherry*-FRT-*ReaChR.* A DNA fragment containing the ReaChR::Citrine coding sequence was amplified by PCR using primers (ReaChR-f and ReaChR-citrine-r; see **Supplementary Table 4**). This PCR product was subcloned into pJFRC2 and pJFRC19 (ref. 30) using SLIC cloning[42] for UAS- and LexAop-driven versions, respectively. For the version containing an *FRT*–mCherry–Stop-*FRT* cassette, the *FRT* sequences (GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTC) and ReaChR DNA fragments were subcloned together into pJFRC2 and pJFRC19 using SLIC cloning[42]. These vectors were injected and integrated into attP40, attP5 and VK5 sites[30].

Fly strains. UAS-ChR2 (ref. 5), UAS-dTrpA1 (ref. 15), UAS-GCaMP3.0 (ref. 40), Gr5a-*GAL4* (ref. 18) and BDP-*GAL4* (ref. 43) (empty promoter *GAL4*: an enhancerless *GAL4* containing a *Drosophila* basal promoter) were generously provided by A. Fiala, P.A. Garrity, L.L. Looger, K. Scott and G.M. Rubin, respectively. *fru-GAL4* (ref. 27), *fru-FLP*[38] and VT40556 *GAL4* (ref. 28) were kindly provided by B.J. Dickson. *hb9-GAL4* was obtained from Bloomington Stock Center (BL #32555). *Crz-GAL4* (ref. 26) and UAS-C128T[12] were previously created in the lab. All the transgenic flies created for this paper are summarized in **Supplementary Table 1**. These flies are available on request.

All experimental flies were maintained on a 12/12 h day-night cycle. Newly eclosed male flies were CO_2 anesthetized and allowed to recover for more than 3–7 d before behavioral tests at 25 °C. For dTrpA1 experiments, flies were raised at 18 °C. For experiments with Gr5a-*GAL4*, female flies were used; for all the other experiments, male flies were used.

Feeding of retinal. All-*trans*-retinal powder (Sigma) was stored in −20 °C as a 40 mM stock solution dissolved in DMSO (×100). 400 µl of sugar-retinal solution (400 µM all-*trans*-retinal diluted in 89 mM sucrose) was directly added to surface of solid food in food vials when larvae were at the first or second instar stage. After collection of newly eclosed flies, they were transferred into a vial containing food with 400 µM all-*trans*-retinal (food was heated and liquefied to mix the retinal evenly in the food). We found that larval feeding is not necessary, but it was performed for all the experiments in this paper to be consistent.

Behavioral setup. See **Supplementary Table 2** for a list of components used to assemble the behavioral setup. See **Supplementary Figure 1** for details of the setup and the behavioral chamber. In brief, high-power LEDs mounted on heat sinks were placed above the behavioral chamber to provide an illumination source (**Fig. 2a** and **Supplementary Fig. 1a**,**b**). The range of available light intensities in our setup is approximately 0.001–1 mW/mm^2 (note that intensity ranges are different for different LEDs; see **Supplementary Fig. 1d**). LED units were designed to be switchable to facilitate testing of different photostimulation wavelengths. The LEDs were controlled by an externally dimmable LED driver (700 mA, externally dimmable, Buckpuck DC driver with leads), and its output was adjusted using custom software controlling an Arduino Uno board (Smart Projects). The Arduino digital PWM output was converted into analog voltage using an RC filter (electronic low-pass filter composed of resistor and capacitor; RC LPF in **Fig. 2a**) containing a 200-Ω resistor and 1-µF capacitor to control the output current of the LED driver. Fly behavior was monitored using a CMOS camera equipped with an IR long-pass filter to avoid detection of light from the high-power LEDs. IR back light was used to visualize the behaving flies. Video capture and LED control were time locked using the Arduino Uno board. To time-stamp photostimulation trials in the videos, we placed an IR indicator LED, whose illumination was synchronized to that of the photostimulation LEDs, in the field of view of the camera. The temperature inside the behavioral chamber was minimally affected by the high-intensity photostimulation: after illumination using the highest available intensities of blue, green or red LEDs (1.1, 0.67 and 1.27 mW/mm^2, respectively) for 1 min, the biggest change in ambient temperature, detected using a thermocouple inserted into the chamber, was 0.7 °C.

Behavioral experiments and quantification of behaviors. For experiments to activate Gr5a-GRNs, nonstarved flies were mounted into 200-µl Pipetman tips as described previously[12]. Mounted flies were placed beneath high-power LEDs, and PERs were monitored using a video camera. Mounted flies were not placed in the behavioral chamber but placed at the same location as the wells of behavioral chamber in **Supplementary Figure 1b**. Bouts of PER were counted manually. A bout was defined as beginning when flies start extending their proboscis and ending when they retract the proboscis. Incomplete proboscis extensions were not counted. LEDs were used at maximum intensities in **Figures 1b**,**c** and **2d**,**e** (red, 1.1 mW/mm^2; amber, 0.22 mW/mm^2; green, 0.67 mW/mm^2; blue, 1.27 mW/mm^2). For **Figure 1b**, 100-ms photostimulation trials (1 Hz) were delivered (three trials), and flies showing more than one PER during this activation period were counted as responders. Fly genotype: *w*$^-$; Gr5a-*GAL4*(II); GR5a-*GAL4*(III)/UAS-ReaChR(VK5) (**Fig. 1b–g**); *w*$^-$; Gr5a-*GAL4*(II)/UAS-dTrpA1(II); GR5a(III)-*GAL4*/UAS-dTrpA1(III) (**Fig. 1h**).

To activate Crz neurons (**Fig. 2d**), males expressing each opsin in *Crz-GAL4* neurons were mounted dorsal side down on a glass

slide as previously described[26]. Flies were illuminated using the maximum available intensity of light for each type of LED, continuously for 1 min, while we monitored them from the ventral side using a video camera. The number of flies exhibiting ejaculation during light stimulation was manually counted.

For all other behavioral experiments, we used acrylic behavioral chambers (16-mm diameter) in a 2 × 4 array (**Fig. 2** and **Supplementary Fig. 1**) to monitor fly behavior. Unless otherwise indicated, chambers were photostimulated using the maximum intensity available for each LED, for 1 min using continuous illumination, while we monitored them with the camera from above. The number of flies showing continuous side walking during stimulation using the *hb9-GAL4* driver was manually counted (**Fig. 2d**,**f**). *fru-GAL4* neurons were activated in the same manner, and flies showing wing extension or paralysis phenotypes were counted manually (**Fig. 2d**,**g**). Paralysis was defined as the cessation of locomotion and loss of postural control. Flies that showed a weaker behavioral phenotypes (HB9, side walk; Fru, wing extension) at the onset of photostimulation, but that were paralyzed before the 1-min stimulation was terminated, were counted as paralysis (**Fig. 2f**,**g**).

Wing extension evoked by activation of P1 or pIP10 neurons were tested in solitary males in the absence of female flies. The wing extension was manually scored (**Figs. 2–5**). Grooming (rapid wing movements while touching with hind leg) was excluded. A bout was defined as starting when flies begin to increase the wing angle and ending when they stop decreasing it.

In order to fit the data into a sigmoidal curve, sigmoid interpolation was performed. The sigmoid curves were defined as follows

$$F_{\text{behav}} = \frac{1}{1 + e^{-\alpha \log_2 \frac{X}{X_{50}}}}$$

where F_{behav} is the fraction of flies showing the behavior, X is the light intensity (**Figs. 3f**, **4c** and **5c**) or frequency (**Fig. 2e**), X_{50} is the light intensity (**Figs. 3f**, **4c** and **5c**) or frequency (**Fig. 2e**) where 50% of flies show the behavior, and α is the slope of the sigmoid curve.

On the basis of the experimentally measured quantities (X and F_{behav}), X_{50} and α were chosen to best fit the data. For all experimental data, polynomial curve fitting—which finds the coefficients that fit the data by the least-squares method—was calculated with Matlab (MathWorks). Goodness of fit was tested by two-way ANOVA between the sigmoidal curve and the actual PER response curve, which indicated a good fit for all cases ($P < 0.05$, two-way ANOVA). The X_{50} is shown as 50% point in the figures (**Figs. 2e**, **3f**, **4c** and **5c**).

Measurement of light intensity. A photodiode power sensor (S130VC, Thorlabs) was placed at the location of the behavioral chamber but in the absence of the chamber. The peak wavelength of each LED (red, 627 nm; amber, 590 nm; green, 530 nm; blue, 470 nm) was measured at different voltage inputs. Measurements were repeated four times and averaged. The baseline intensity of each wavelength before LED illumination was subtracted. Note that light intensity can drop during stimulation at high input voltages. In this study, intensity after 10 s of stimulation was measured.

Measurement of penetrance of different wavelengths of light through the fly cuticle. The proboscis of a female adult fly was removed, and a 10-μm multimode optic fiber (NA, 0.1; Thorlabs) was inserted into the brain through the window. The amount of light entering the optic fiber inside or outside the fly was measured using a power meter (Model 1931, Newport). Penetrance was calculated as the amount of light that entered the optic fiber inside the fly divided by the amount of light measured outside the fly. The long axis of the optic fiber was always aligned with the light source. Different wavelengths of high power LEDs (470 nm, 530 nm, 590 m, 627 nm) were used as light sources.

Fly histology. All fixation and staining procedures were performed at 4 °C in PBS unless otherwise specified. Dissected brains were fixed in 4% formaldehyde in PEM (0.1 M PIPES, pH 6.95, 2 mM EGTA, 1 mM $MgSO_4$) for 2 h. After three 15-min rinses with PBS, brains were incubated with primary antibodies overnight. Following three 15 min rinses with PBS, brains were incubated with secondary antibody overnight. Following three 15-min rinses, brains were incubated in 50% glycerol in PBS for 2 h and cleared with Vectashield (Vector Labs). All procedures were performed at 4 °C. A FluoView FV1000 Confocal laser scanning biological microscope (Olympus) with a 30×/1.05-NA silicone oil objective (Olympus) was used to obtain confocal serial optical sections. The antibodies used for **Supplementary Figure 2a**,**d** were anti-GFP, rabbit polyclonal antibody unconjugated (A11122, Invitrogen) and Alexa Fluor 488 donkey anti-rabbit IgG(H+L) (A11008, Invitrogen). Both of the antibodies were diluted to 1/300. Expression of mCherry in **Supplementary Figure 2d** was monitored using native fluorescence without antibody staining.

FluoRender software[44] (http://www.sci.utah.edu/software/13-software/127-fluorender.html) was used to make 3D image reconstructions. To measure the expression levels of ReaChR::Citrine in P1 neurons in **Figure 6d**, the native fluorescence of Citrine in different specimens was monitored using the same intensity of laser power (470 nm) and PMT voltage. Signal intensity was quantified in ImageJ (http://rsbweb.nih.gov/ij/).

Calcium imaging. Two-photon imaging was performed on an Ultima two-photon laser-scanning microscope (Prairie Technology) with an imaging wavelength of 925 nm (**Fig. 6**). To filter out autofluorescence of the brain and light from the amber stimulation LED (for ReaChR activation), we used a 500/20 nm (center wavelength/bandwidth) band-pass filter (Chroma) in the emission pathway to detect the GCaMP3 fluorescence. With this laser and filter setting, fluorescence emissions from the Citrine tag (on ReaChR) were not detectable by our PMT. This was confirmed by examination of P1-GAL4;UAS-ReaChR::Citrine flies (the flies without GCaMP3.0), which exhibited no fluorescence signal under our imaging conditions. Therefore, the detected fluorescence signals are purely from GCaMP3.0. The scanning resolution was 128 × 128 pixels, dwell time per pixel was 8 μsec and the optical zoom was 4×. The scanning speed was ~10 Hz. The excitation intensity of the two-photon laser was varied among samples depending on the level of GCaMP3.

doi:10.1038/nmeth.2765

In both cases, a 40×/0.80-NA water-immersion objective (Olympus) was used for imaging. A high-power amber LED (590 nm) collimated with an optic fiber (M590F1, Thorlabs) was used as a light source to activate ReaChR. To narrow the bandwidth of the LED output, we connected the optic fiber to a fiber optic filter holder (World Precision Instruments) equipped with 589/10 nm (center wavelength/bandwidth) band-pass filter (Edmund optics). A 200-µm core multimode optic fiber (NA, 0.39; FT200EMT, Thorlabs) was used to deliver the light from the fiber optic holder to the brain. One side of the optic fiber was custom-made to be a bare tip (Thorlabs) and was dipped into the saline imaging bath and placed 430 µm away from the brain. A 10×/0.30-NA water-immersion objective (Olympus) was used to locate the brain and align the optic fiber. The distance between brain and the fiber was measured with an objective micrometer (Olympus). We set the light intensity to be 170 µW at the tip of optic fiber. Thus, at a distance of 430 µm from the tip of a 0.39-NA optic fiber, the light power is calculated to be approximately 1.7 mW/mm^2 at the brain surface (the size of the light spot should be approximately 0.10 mm^2 at the brain). In addition to the PMT used to monitor GCaMP emissions, we used another PMT to monitor the 590-nm ReaChR activation light. This was to ensure that the intensities of 590-nm light were comparable between samples.

To prepare the brain for imaging, we used an *ex vivo* prep. After briefly anesthetizing a fly on ice, the brain was dissected out using a sharp forceps into a 35-mm plastic Petri dish (35 3001, Falcon) containing *Drosophila* imaging saline (108 mM NaCl, 5 mM KCl, 2 mM $CaCl_2$, 8.2 mM $MgCl_2$, 4 mM $NaHCO_3$, 1 mM NaH_2PO_4, 5 mM trehalose, 10 mM sucrose, 5 mM HEPES, pH 7.5)[45]. The fat body, air sacs and esophagus were gently removed to give a clear view of the brain and to minimize its movement. The brains were attached to the bottom of the plate by static. The saline was changed once after dissection to remove debris. Calcium imaging was performed within 10–15 min after the dissection to ensure that the brains were healthy.

Electrophysiology. The tip recording method was used to record the electrophysiological responses of labellar taste neurons[46]. Briefly, the fly was mounted and immobilized for recording by inserting a pulled glass capillary (BF150-86-10, Sutter Instruments) from the dorsal surface of the thorax to the tip of the labellum, passing through the cervical connective and the head. The mounting glass capillary was filled with recording solution (7.5 g/L NaCl, 0.35 g/L KCl, 0.279 g/L $CaCl_2{\cdot}2H_2O$ and 11.915 g/L HEPES (Sigma-Aldrich)) and served as a ground electrode. Another glass capillary, pulled to a tip diameter of 10–20 µm and filled with 30 mM tricholine citrate (TCC; Sigma-Aldrich), as an electrolyte, was used for recording the electrophysiological responses of the gustatory neurons innervating this sensillum. All the recordings were obtained from L7 sensilla. The recordings were made using a MultiClamp 700B amplifier and Digidata 1440A A/D converter (Molecular Devices). The recorded data were sampled at a rate of 10 kHz, filtered (band-pass filter between 100 Hz and 3 kHz) and stored on a PC hard drive using Clampex 10 software (Molecular Devices). The data were analyzed by sorting the action potentials and measuring their frequency within the indicated time windows using Clampfit software (Molecular Devices).

For PER activation experiments, a high-power amber LED (590 nm) collimated with an optic fiber (M590F1, Thorlabs) was used as a light source to activate ReaChR. For delivering light to the labellum, a 200-µm core multimode optic fiber with bare end (NA, 0.39; Thorlabs) was used. The distance of optic fiber from the labellum was set to be 540 µm using a micrometer. The estimated light intensity at the labellum was approximately 1.0 mW/mm^2.

To activate TrpA1 (**Fig. 1h**), we used a custom-made heat source. In brief, the heat source is a small piece of thermistor (2K Bead Thermistor, Fenwal) that emits heat in proportion to the electrical current passed through it. The distance of the heat source from the labellum was set to be 540 µm using a micrometer. The temperature at this distance was measured using a thermocouple (Omega) (top panel in **Fig. 1h**).

41. Donnelly, M.L. *et al.* The 'cleavage' activities of foot-and-mouth disease virus 2A site-directed mutants and naturally occurring '2A-like' sequences. *J. Gen. Virol.* **82**, 1027–1041 (2001).
42. Li, M.Z. & Elledge, S.J. Harnessing homologous recombination *in vitro* to generate recombinant DNA via SLIC. *Nat. Methods* **4**, 251–256 (2007).
43. Pfeiffer, B.D. *et al.* Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **105**, 9715–9720 (2008).
44. Wan, Y., Otsuna, H., Chien, C.B. & Hansen, C. An interactive visualization tool for multi-channel confocal microscopy data in neurobiology research. *IEEE Trans. Vis. Comput. Graph.* **15**, 1489–1496 (2009).
45. Wong, A.M., Wang, J.W. & Axel, R. Spatial representation of the glomerular map in the *Drosophila* protocerebrum. *Cell* **109**, 229–241 (2002).
46. Hodgson, E.S., Lettvin, J.Y. & Roeder, K.D. Physiology of a primary chemoreceptor unit. *Science* **122**, 417–418 (1955).

doi:10.1038/nmeth.2765

ONLINE METHODS

Experimental details. We used data from five different cancer types available from the TCGA website: GBM, BIC, LSCC, KRCCC and COAD. For each of these tumor types, we downloaded TCGA-curated level 3 data sets containing gene expression, miRNA expression and DNA methylation information. TCGA repository contains multiple platforms for each data type. We always chose the platform corresponding to the largest number of available individuals and describing both tumor samples and controls whenever possible. For expression data, we used the Broad Institute HT-HG-U133A platform in GBM and LSCC, the UNC-Agilent-G4502A-07 platform in BIC and COAD and the UNC-Illumina-Hiseq-RNASeq platform in KRCCC. For miRNA expression data, we used the BCGSC-Illumina-Hiseq-miRNAseq platform in BIC, the UNC-miRNA-8X15K platform in GBM and the BCGSC-Illumina-GA-miRNAseq in LSCC, KRCCC and COAD. Finally, for the methylation data we used the JHU-USC-Illumina-DNA-Methylation platform in GBM, the JHU-USC-Human-Methylation-27 platform for BIC, LSCC, KRCCC and COAD. For all these tumor types, we also downloaded patients' clinical information including the overall survival data.

We also used METABRIC data set to evaluate the effectiveness of survival prediction with network regularization. METABRIC data set consists of two cohorts: discovery (997 patients) and validation (995 patients). For each of these patients, matched DNA and RNA were extracted from each primary tumor specimen and subjected to copy-number and genotype analysis on the Affymetrix SNP 6.0 platform and transcriptional profiling on the Illumina HT-12 v3 platform (Illumina-Human-WG-v3). We used the normalized data available from the European Genome-Phenome Archive (EGA, http://www.ebi.ac.uk/ega/). High-quality follow up clinical data including information on disease-free survival were also available for both cohorts. As a preprocessing step, we mapped copy-number variations to genes using the PennCNV package[20].

Before applying our SNF, we performed three steps of preprocessing: outlier removal, missing-data imputation and normalization. If a patient had more than 20% missing data in a certain data type, we did not consider this patient. Similarly, if a certain biological feature (for example, mRNA expression) had more than 20% of missing values across patients, we filtered out this feature. Also, for missing data, we used K nearest neighbor (KNN) imputation[21], where the number of neighbors is the same with K value used in our method (see below); therefore we do not have any free parameters. Last, before constructing the patient network, we performed the following normalization:

$$\tilde{f} = \frac{f - E(f)}{\sqrt{\mathrm{Var}(f)}},$$

where f is any biological feature, $\tilde{f}$ is the corresponding feature after normalization, $E(f)$ and $Var(f)$ represent the empirical mean and variance of f, respectively.

Evaluation metrics. We used several metrics for evaluation and comparison of our method to existing approaches. In the real-cancer data, we use three metrics, as ground truth was not known. First, we use silhouette[15] to measure the homogeneity of the subtypes. For each patient i, let $a(i)$ denote the average dissimilarity to all other patients within the same subtype and $b(i)$ denote the lowest average dissimilarity to all other patients in different subtypes. The value of silhouette for patient i was defined as $s(i) = (b(i) - a(i))/(\max a(i), b(i))$. The mean value of silhouette for all the patients was then used as a measure of how tightly grouped all the data in the cluster are. If silhouette value was close to 1, then it means the data were appropriately clustered.

We also used P value for log-rank test of survival separation in Cox regression model[14]. P value measures the significance in the difference of survival profiles between subtypes. In our test, we set 0.05 to be the threshold of the significance. The lower the P value was, the less likely it was that such differential survival was observed by chance, i.e., the more significantly different the survival profiles was between subtypes. For most cancers, we used days to the last follow-up and the vital status to perform the log-rank test for survival analysis. However, for COAD, we used the consensus of the days to last known alive together with the last follow-up as a proxy because there were a lot of missing values in the data for days to last follow up. We used running time (in minutes) to compare the scalability of each method.

Similarity network fusion. Suppose we have n samples (for example, patients) and m measurements (for example, mRNA gene expression). We will use the patient network example throughout this section for clarity though the method has broad applicability as discussed above. A patient similarity network is represented as a graph $G = (V, E)$. The vertices V correspond to the patients $\{x_1, x_2, \cdots, x_n\}$ and the edges E are weighted by how similar the patients are. Edge weights are represented by an $n \times n$ similarity matrix $\mathbf{W}$ with $\mathbf{W}(i, j)$ indicating the similarity between patients x_i and x_j and are computed as follows. We denote $\rho(x_i, x_j)$ as the Euclidean distance between patients x_i and x_j. We then use a scaled exponential similarity kernel to determine the weight of the edge:

$$\mathbf{W}(i, j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu\varepsilon_{i,j}}\right) \quad (1)$$

where μ is a hyperparameter that can be empirically set and $\varepsilon_{i,j}$ is used to eliminate the scaling problem. Here we define

$$\varepsilon_{i,j} = \frac{\mathrm{mean}(\rho(x_i, N_i)) + \mathrm{mean}(\rho(x_j, N_j)) + \rho(x_i, x_j)}{3}$$

where mean $(\rho(x_i N_i))$ is the average value of the distances between x_i and each of its neighbors. We recommend setting μ in the range of [0.3, 0.8]. Note that while this distance measure is suitable for continuous variables, we propose to use chi-squared distance for discrete variables and agreement-based measure for binary variables.

To compute the fused matrix from multiple types of measurements, we define a full and sparse kernel on the vertex set V. The full kernel is a normalized weight matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}$ is the diagonal matrix whose entries $\mathbf{D}(i, i) = \Sigma_j \mathbf{W}(i, j)$, so that $\Sigma_j P(i, j) = 1$. However, this normalization may suffer from numerical instability since it involves self-similarities on the diagonal entries of $\mathbf{W}$. One way to perform a better normalization is as follows:

$$\mathbf{P}(i, j) = \begin{cases} \dfrac{\mathbf{W}(i, j)}{2\Sigma_{k \neq i}\mathbf{W}(i,k)}, & j \neq i \\ 1/2, & j = i \end{cases} \quad (2)$$

doi:10.1038/nmeth.2810

This normalization will be free of the scale of self-similarity in the diagonal entries and $\Sigma_j \mathbf{P}(i,j) = 1$ still holds.

Let N_i represent a set of x_i's neighbors including x_i in G. Given a graph, G, we use K nearest neighbors (KNN) to measure local affinity as:

$$\mathbf{S}(i,j) = \begin{cases} \dfrac{\mathbf{W}(i,j)}{\Sigma_{k \in N_i} \mathbf{W}(i,k)}, & j \in N_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

This operation sets the similarities between non-neighboring points (in terms of the pairwise similarity values) to zero. Essentially we make the assumption that local similarities (high values) are more reliable than remote ones; and we thus assign similarities to non-neighbors through graph diffusion on the network. This is a mild assumption widely adopted by other manifold learning algorithms. Note that $\mathbf{P}$ carries the full information about the similarity of each patient to all others whereas $\mathbf{S}$ only encodes the similarity to the K most similar patients for each patient. Our algorithm always starts from $\mathbf{P}$ as the initial state using $\mathbf{S}$ as the kernel matrix in the fusion process for both capacity of capturing local structure of graphs and computational efficiency.

Given m different data types, we can construct similarity matrices $\mathbf{W}^{(v)}$ using equation (1) for the v^{th} view, $v = 1,2,\ldots,$ m. $P^{(v)}$ and $\mathbf{S}^{(v)}$ are obtained from equations (2) and (3), respectively. Below we introduce our network fusion process given a set of networks.

Let us first consider the case when we have two data types, i.e., $m = 2$. We calculate the status matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ as in equation (2) from two input similarity matrices; then the kernel matrices $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are obtained as in equation (3).

Let $\mathbf{P}^{(1)}_{t=0} = \mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}_{t=0} = \mathbf{P}^{(2)}$ represent the initial two status matrices at $t = 0$. The key step of SNF is to iteratively update similarity matrix corresponding to each of the data types as follows:

$$\mathbf{P}^{(1)}_{t+1} = \mathbf{S}^{(1)} \times \mathbf{P}^{(2)}_t \times (\mathbf{S}^{(1)})^T \quad (4)$$

$$\mathbf{P}^{(2)}_{t+1} = \mathbf{S}^{(2)} \times \mathbf{P}^{(1)}_t \times (\mathbf{S}^{(2)})^T \quad (5)$$

where $\mathbf{P}^{(1)}_{t+1}$ is the status matrix of the first data type after t iterations. $\mathbf{P}^{(2)}_{t+1}$ is the similarity matrix for the second data type. This procedure updates the status matrices each time generating two parallel interchanging diffusion processes. After t steps, the overall status matrix is computed as

$$\mathbf{P}^{(c)} = \frac{\mathbf{P}^{(1)}_t + \mathbf{P}^{(2)}_t}{2}.$$

Since $\mathbf{S}$ is a KNN graph of $\mathbf{P}$, which can reduce some noise between instances, our SNF is robust to the noise in similarity measures.

Another way to think of the updating rule (4) is

$$\mathbf{P}^{(1)}_{t+1}(i,j) = \sum_{k \in N_i} \sum_{l \in N_j} \mathbf{S}^{(1)}(i,k) \times \mathbf{S}^{(1)}(j,l) \times \mathbf{P}^{(2)}_t(k,l) \quad (6)$$

(the same for $\mathbf{P}^{(2)}_{t+1}$). Note N_i represents the neighborhood of x_i. We can see that similarity information is only propagated through the common neighborhood. This renders SNF robust to noise. An important observation is that if x_i and x_j have common neighbors in both similarity matrices, it is highly possible that they belong to the same cluster. Another essential fact our method benefits from is that even if x_i and x_j are not very similar in one data type, their similarity can be expressed in another data type and this similarity information can be propagated through the fusion process.

After each iteration, we performed normalization on $\mathbf{P}^{(1)}_{t+1}$ and $\mathbf{P}^{(2)}_{t+1}$ as in equation (2). By performing the normalization, we (i) ensure that throughout SNF iterations a patient is always most similar to himself than to other patients; (ii) ensure that our final network is full rank, important for the classification and clustering applications of the final network. Finally, we have found that the use of such normalization leads to quicker convergence of SNF.

Finally, an extension to the case $m > 2$ follows equations (4) and (5):

$$\mathbf{P}^{(v)} = \mathbf{S}^{(v)} \times \left(\frac{\Sigma_{k \neq v} \mathbf{P}^{(k)}}{m-1} \right) \times (\mathbf{S}^{(v)})^T, v = 1,2,\cdots,m \quad (7)$$

The input to our algorithm can be feature vectors, pairwise distances, or pairwise similarities. The learned status matrix $P^{(c)}$ can then be used for retrieval, clustering and classification; in this work, we focus mostly on clustering and prediction.

SNF is inspired by the theoretical multiview learning framework developed for the computer vision and image processing applications[22] that is not directly applicable to biological data. SNF constructs networks of samples (for example, patients) by comparing samples' molecular (or phenotypic) profiles; fused networks are used for subtyping and label prediction distinguishing SNF from all the previously published research.

Network clustering (for example, for disease subtyping). Given n samples and m measurements we want to identify C clusters of samples, each of which corresponds to a (known or new) subtype. We associate each sample x_i with a label indicator vector $\boldsymbol{y}_i \in \{0,1\}^C$ such that $\boldsymbol{y}_i(k) = 1$ if sample x_i belongs to the k^{th} cluster (subtype), otherwise $\boldsymbol{y}_i(k) = 0$. So a partition matrix $\mathbf{Y} = \left(\boldsymbol{y}_1^T; \boldsymbol{y}_2^T; \cdots; \boldsymbol{y}_n^T\right)$ is used to represent a clustering scheme.

Given the fused graph, in this work we used spectral clustering to obtain network clusters. Traditional state-of-the-art spectral methods[23], aim to minimize RatioCut[24], an objective function that effectively combines MinCut and equipartitioning, by solving the following optimization problem:

$$\min_{Q \in \mathbb{R}^{n \times C}} \text{Trace}(\mathbf{Q}^T \mathbf{L}^+ \mathbf{Q}) \quad \text{s.t.} \mathbf{Q}^T \mathbf{Q} = \mathbf{I} \quad (8)$$

where $\mathbf{Q} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}$ is a scaled partition matrix, $\mathbf{L}^+$ denotes the normalized Laplacian matrix $\mathbf{L}^+ = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ given the similarity matrix $\mathbf{W}$). Matrix $\mathbf{D}$ is a network degree matrix, with degrees of each node on the diagonal and off-diagonal elements set to 0). Spectral clustering is effective in capturing global structure of the graph[25].

Network-based survival risk prediction. With the fused network, we can perform tasks beyond disease subtyping. An example in this paper is survival prediction with network regularization. Cox model has been successfully applied to perform survival/risk prediction of given new patients. Given all the feature matrix $\mathbf{X}$, the risk of an event (death) at time t for the i-th patient is given by $h(t|\mathbf{X}) = h_0(t) \exp(\mathbf{X}^T z)$, where z is a vector of regression coefficients and $h_0(t)$ is the baseline hazard function. This regression

coefficient vector $\boldsymbol{z}$ is estimated by maximizing the Cox's log-partial likelihood:

$$lp(z) = \sum_{i=1}^{n} \delta_i \left(\mathbf{X}_i^T z - \log\left(\sum_{j \in R(t_i)} \exp\left(\mathbf{X}_j^T z \right) \right) \right) \quad (9)$$

where n is the number of patients, t_i is the survival time for the i-th patient and $R(t_i)$ is the risk set at time t_i, i.e., the set of patients who still survived before t_i. $\delta_i(\cdot)$ is an indicator function whether the survival time is observed ($\delta_i = 1$) or censored ($\delta_i = 0$).

It is possible to improve survival prediction by incorporating additional information, such as gene interaction data[26] or patient similarity based constraints. To incorporate the network structure, similarity between either features or patients (or both) can be used as a regularizer. According to the hazard function of Cox's model, the relative risk between patient i and patient j is $\exp(\mathbf{X}_i^T z - \mathbf{X}_j^T z)$, therefore, a regularizer can be constructed as $(\mathbf{X}_i^T z - \mathbf{X}_j^T z)^2 w_{ij}$. To estimate $\boldsymbol{z}$, we can use a modified likelihood expression as follows:

$$lp(z) = \sum_{i=1}^{n} \delta_i \left(X_i^T z - \log\left(\sum_{j \in R(t_i)} \exp(X_j^T z) \right) \right) - \lambda \sum_i \sum_j (X_i^T z - X_j^T z)^2 w_{ij} \quad (10)$$

where λ is the regularizing coefficient. Newton optimization techniques are applied to solve this maximization problem.

Combining data types. SNF can be used to incorporate arbitrary types of discrete (binary or categorical) and continuous data. For integration of discrete data, we recommend the use of chi-squared distance as the similarity measure. Compatibility of data sources can be checked via normalized mutual information (NMI). If the patient similarity obtained from different data sources is completely discordant; NMI can help to clarify which data should and which should not be combined.

20. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
21. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
22. Wang, B., Jiang, J., Wang, W., Zhou, Z.-H. & Tu, Z. Unsupervised metric fusion by cross diffusion. in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2997–3004 (IEEE, 2012).
23. Ng, A.Y., Jordan, M.I. & Weiss, Y. On spectral clustering: analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2**, 849–856 (2002).
24. Wei, Y.C. & Cheng, C.K. Towards efficient hierarchical designs by ratio cut partitioning. in *Proc. Int. Conf. Computer-Aided Design* 298–301 (ICCAD, 1989).
25. Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
26. Zhang, W. *et al.* Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput. Biol.* **9**, e1002975 (2013).

doi:10.1038/nmeth.2810