doi:10.1038/nature05316

LETTERS

Variability and memory of protein levels in human cells

Alex Sigal[1]*, Ron Milo[1]*†, Ariel Cohen[1]*, Naama Geva-Zatorsky[1], Yael Klein[1], Yuvalal Liron[1], Nitzan Rosenfeld[1], Tamar Danon[1], Natalie Perzov[1] & Uri Alon[1]

Protein expression is a stochastic process that leads to phenotypic variation among cells[1–6]. The cell–cell distribution of protein levels in microorganisms has been well characterized[7–23] but little is known about such variability in human cells. Here, we studied the variability of protein levels in human cells, as well as the temporal dynamics of this variability, and addressed whether cells with higher than average protein levels eventually have lower than average levels, and if so, over what timescale does this mixing occur. We measured fluctuations over time in the levels of 20 endogenous proteins in living human cells, tagged by the gene for yellow fluorescent protein at their chromosomal loci[24]. We found variability with a standard deviation that ranged, for different proteins, from about 15% to 30% of the mean. Mixing between high and low levels occurred for all proteins, but the mixing time was longer than two cell generations (more than 40 h) for many proteins. We also tagged pairs of proteins with two colours, and found that the levels of proteins in the same biological pathway were far more correlated than those of proteins in different pathways. The persistent memory for protein levels that we found might underlie individuality in cell behaviour and could set a timescale needed for signals to affect fully every member of a cell population.

We asked whether and on what timescale do protein levels mix in individual human cells. Mixing, in the present context, occurs when a cell lineage, given enough time, reaches the different states found in a snapshot of a cell population. Irreversible differences between cells, such as expected from genetic changes, would lead to cells that retain their protein state indefinitely without mixing (Fig. 1a, left). Mixing behaviour (Fig. 1a, middle and right) can be characterized by a typical timescale τ_m, called the mixing time, after which cells lose the memory of their previous protein state. Recent studies using synthetic protein expression systems in bacteria indicate that cells with

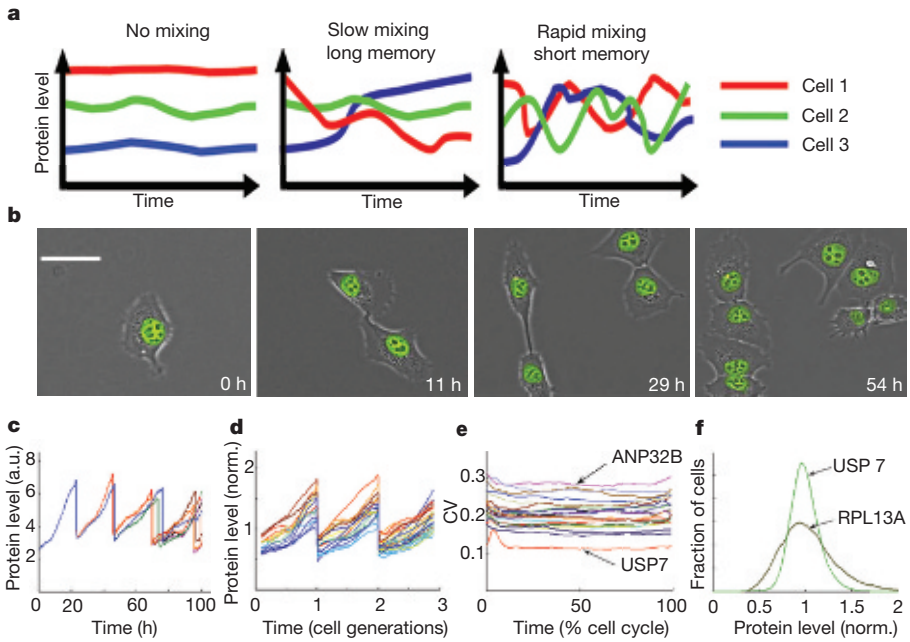

Figure 1 | Protein level dynamics in individual cells. **a**, Possible mixing dynamics of a protein in a population of cells. **b**, Cells expressing YFP CD-tagged topoisomerase 1 (TOP1) at different time-points. All cells are the progeny of the cell shown in the first frame. Scale bar is 25 μm **c**, TOP1 levels as a function of time in a cell lineage. Sharp decreases in protein levels indicate cell-division events and lines originating after each division are the TOP1 levels of daughter cells. **d**, TOP1 protein dynamics synchronized to a cell-cycle time-base using cell division events. Protein levels were normalized by the average protein level. **e**, The coefficient of variation (CV = s.d./mean) across the cell cycle of 20 different proteins. **f**, Distribution of protein levels measured by flow cytometry for cells at a similar cell-cycle stage, for USP7 (CV = 0.16) and RPL13A (CV = 0.28).

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, 76100 Israel. †Present address: Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

high production tend to maintain this for about a cell generation[12], yielding mixing times on the order of 1–2 h (ref. 13). However, little quantitative information exists about variability in protein levels among human cells, and specifically among human cancer cells, where acquired genetic differences might add to non-genetic sources of noise.

To examine the extent and persistence of variability in the levels of different proteins, we used a fluorescent labelling strategy known as CD-tagging, which tags proteins at their endogenous chromosomal locations[24–26]. A yellow fluorescent protein (*YFP*) sequence was retrovirally inserted into the genome of human H1299 lung carcinoma cells. The *YFP* sequence contained splice acceptor and donor flanking sequences, which enabled the tag to be spliced into the mRNA of a gene as a new exon, resulting in the expression of a fusion protein[24,25]. The tagged gene was therefore expressed under its native promoter on the chromosome. CD-tagging has been shown to preserve the localization and function of most of the tagged proteins[24–27].

The dynamics of YFP-tagged proteins in individual cells were followed by time-lapse fluorescence microscopy under controlled temperature, humidity and CO_2 conditions (Fig. 1b, see also Supplementary Movie 1 (transmitted light) and 2 (fluorescence) of cells shown in Fig. 1b). Time-lapse movies were taken over 50–100 h, at a resolution of 1 frame per 10 min.

We quantified protein levels in individual cells over multiple cell cycles (Fig. 1c). We used proteins that were localized in the nucleus to facilitate accurate automatic image analysis of the movies (see Methods). The twenty proteins studied were in a variety of pathways, including RNA splicing, transcriptional regulation, apoptosis and DNA replication (see Supplementary Information). All proteins studied were correctly localized.

To quantify the inherent variability between cells, we synchronized cells *in silico* by aligning protein dynamics between division events[24]. *In silico* synchronization enabled us to examine variability in cells in which an equal fraction of the cell cycle has elapsed, eliminating differences in cell-cycle stage[7,18] as a source of variability (Fig. 1d). We found that the average coefficient of variation of the total fluorescence (CV = standard deviation/mean) depended on the protein measured and ranged from 0.12 to 0.28 (Fig. 1e and Supplementary Information). Protein level distributions were single-peaked and skewed to the right (Fig. 1f). The CV values were nearly constant throughout the cell cycle and were comparable to those found in bacteria and yeast (0.1–1.0)[7,11,14,20]. The ratio between the fluorescence of cells at the 90th percentile to cells at the 10th percentile ranged from about 1.3 to 2.1 (Supplementary Information). The density of cells in our experiments, as measured by the number of cell–cell contacts, did not significantly contribute to the variability in the levels of the proteins examined (see Supplementary Information).

We next studied the mixing dynamics of these proteins. For each protein, we followed a clonal population of cells over several generations using time-lapse microscopy, and synchronized cells to the cell-cycle time base (Fig. 2a, b). For some proteins, such as the ubiquitin-specific protease USP7, mixing occurred in about one cell cycle (~20 h; Fig. 2c). By contrast, other proteins, such as the transcription-regulating factor HMGA2, showed slower mixing, on the timescale of several cell generations (Fig. 2d). Cell division events did not seem to have a strong randomizing effect on protein levels (Fig. 2c, d), in contrast to observations on mRNA levels in bacteria[28].

To compute the mixing time, we calculated the auto-correlation function $A(\tau)$ of the protein levels (Fig. 2e, f and see Methods). The mixing time τ_m was defined as the time when $A(\tau)$ decayed to one half. We found that τ_m ranged from about 0.8 generations for CIRBP, a protein involved in the cold-shock response, to about 2.5 generations for HMGA2 (see Supplementary Information). Furthermore, there was a substantial correlation ($R^2 \approx 0.6$) between the cell–cell variability in protein levels (CV) and the mixing time (Fig. 3). This correlation between mixing time and variability might pose constraints on possible mechanisms for protein noise (see Supplementary Information).

We also investigated the correlations between different proteins in the same cell. For this, we used CD-tagging to generate clones in which two proteins are tagged, one with yellow (*YFP*) and the other with a red (*mCherry*[29]) fluorescent tag. The levels of randomly chosen pairs of proteins from different pathways were nearly uncorrelated

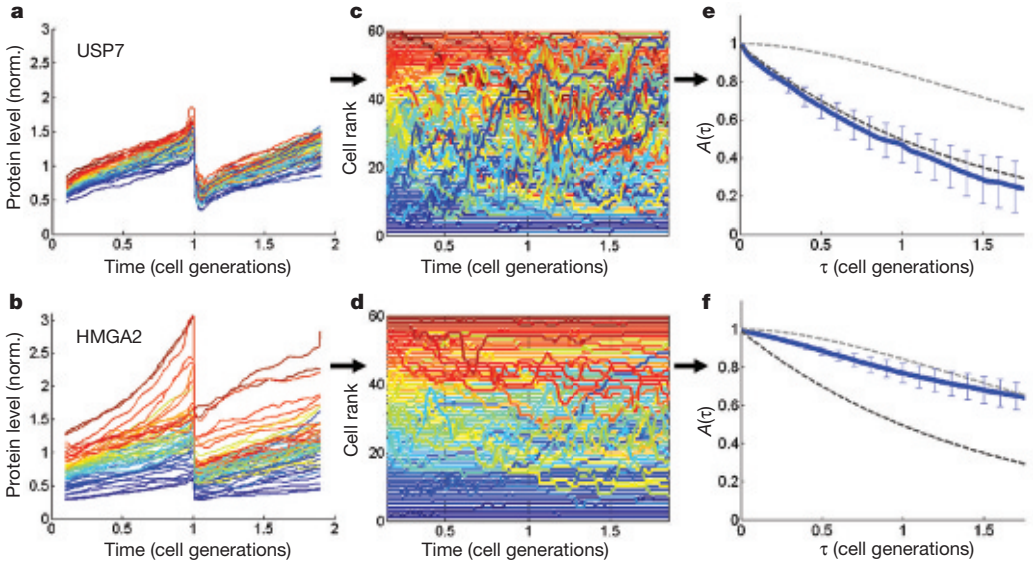

Figure 2 | Mixing-times of different proteins. **a**, **b**, Dynamics of cells expressing YFP CD-tagged USP7 and HMGA2 across two cell generations. Each line represents the protein level of one cell, normalized to the population mean. **c**, **d**, Ranks of 60 randomly selected cells for USP7 and HMGA2 through two cell generations, where 1 is the lowest and 60 is the highest protein level in the population at the given time-point. Lines are colour-coded for rank at the start of the first cell cycle, with red being high relative to the mean and blue low. **e**, **f**, The auto-correlations of protein level ranks, calculated on the basis of at least 150 cells for each protein. The lower dashed line is the theoretical auto-correlation for stable proteins with noisy production, and dilution by cell growth ($A(\tau) = e^{-\alpha\cdot\tau}$, where α is the growth rate). The upper dashed line is the theoretical auto-correlation for proteins whose production rate depends on an upstream stable protein with noisy production, $A(\tau) = e^{-\alpha\cdot\tau}\cdot(1+\alpha\cdot\tau)$. For theoretical models, see Supplementary Information. Bars denote standard errors (bootstrap method).

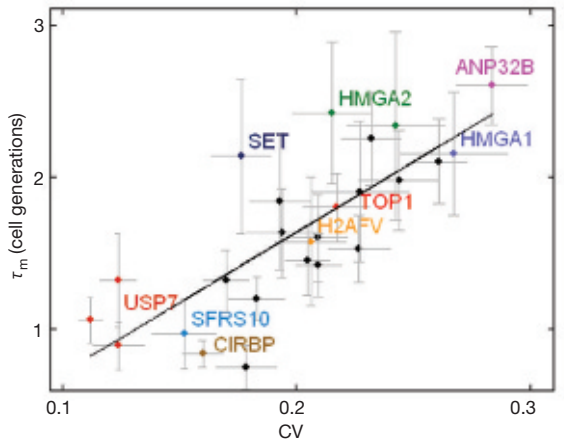

Figure 3 | Mixing time correlates with variability. Mixing time (τ_m) versus CV (s.d./mean) for the 20 proteins described in Supplementary Table 1. The best fit line is shown ($R^2 = 0.62$). Independently derived clones of USP7 and HMGA2 are in red and green. Bars denote standard errors (bootstrap method).

(Fig. 4a, d; see Supplementary Information for functions of proteins studied). The levels of proteins from the same system (ribosome) or levels of proteins produced from different alleles of the same gene (ribosomal protein RPL5) were substantially correlated (Fig. 4b–d). These findings indicate that the main source of noise might be variability in upstream regulatory components specific to each system. Such noise accounted for about 80% of the variance ($\eta^2_{ext}/\eta^2_{tot}$) in levels of RPL5, consistent with measurements of high abundance proteins in microorganisms[7–9,14].

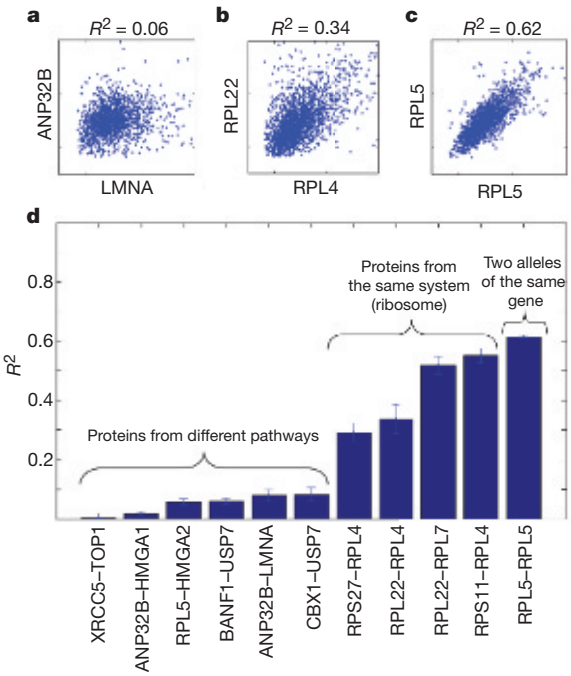

Figure 4 | Correlations between pairs of proteins CD-tagged with two fluorescent colours (YFP and mCherry) in the same cells. Shown is mean nuclear fluorescence of each protein normalized to its population average. **a**, Two proteins with different biological functions: LMNA (component of nuclear membrane) and ANP32B (involved in apoptosis). **b**, Two proteins in the same biological system (two ribosomal proteins, RPL4 and RPL22). **c**, Products of different alleles of the same gene, *RPL5*. The extrinsic and intrinsic noise components[7] of RPL5 are $\eta_{ext} = 0.28$ and $\eta_{int} = 0.14$, and $\eta_{tot} = 0.31$. **d**, All protein pairs measured and their correlations. Bars denote mean $R^2 \pm$ s.e. from three experiments performed on different days.

We compared our finding of long mixing times to simple theoretical models in which proteins are produced and degraded stochastically and diluted out by cell growth (see Supplementary Information). Theoretical analysis leads to the following qualitative features: (1) simple production–degradation models can give mixing times that are at most one cell generation; (2) active protein degradation can reduce the mixing time; and (3) mixing times longer than a cell generation can occur as a result of propagation of noise from upstream components, as suggested by the above correlations observed in proteins from the same pathway, or due to feedback loops.

We found that cell–cell variability in human protein levels can have a long mixing time: proteins whose levels are higher or lower than average can remain so for several days. Such long-memory noise can act, through the cells' regulatory circuitry[3,11,21], to generate variability in cellular responses and phenotypes. In particular, proteins from co-regulated systems such as the ribosome seem to fluctuate in a much more correlated fashion than proteins from different pathways. This long-lasting, concerted noise might result, for example, in 'outlier' cells that have different behaviour from the bulk of the population. The observed long memory indicates that several cell cycles might be needed for some signals and drugs to affect fully populations of human cancer cells.

METHODS

YFP CD-tagging of endogenous proteins. A library of CD-tagged proteins in the H1299 non-small cell lung carcinoma cell line was constructed as described[24]. Briefly, *EYFP*, flanked with splice acceptor and donor sequences, was integrated into the genome using pBabeAE retroviral vectors, each containing *EYFP* in one of three reading frames. Cells positive for YFP fluorescence were sorted using flow cytometry into 384-well plates and grown into cell clones. This screen was gated to preferentially include highly expressed proteins. Tagged protein identities were determined by 3′-RACE (rapid amplification of cDNA ends), using a nested polymerase chain reaction (PCR) that amplified the section between *EYFP* and the polyA tail of the mRNA of the host gene. PCR products were sequenced directly and aligned to the genome. Cyclohexamide treatment of three clones (expressing CD-tagged USP7, HMGA2 and TOP1) indicated that the proteins retained their long lifetime (>24 h) with YFP tagging (not shown). Our library of CD-tagged proteins is detailed at www.dynamicproteomics.net.

Flow cytometry. Cells from exon-tagged clones were labelled with the vital stain SYTO-62 (Molecular Probes), which served as an indication of cell-cycle stage. Fluorescence from each cell was detected at 535 nm (YFP) and 630 nm (SYTO-62) using a Becton Dickinson FACS SORT flow cytometer (BD Biosciences). Cells at a similar cell-cycle stage (bin size of about 5% of the population by SYTO-62 fluorescence) were used for Fig. 1f.

Time-lapse microscopy. We randomly chose 20 nuclear proteins from our CD-tagged library to study variability dynamics. Time-lapse movies were obtained at 20× magnification as described[24] with an automated, incubated Leica DMIRE2 inverted fluorescence microscope and an ORCA ER cooled CCD camera (Hamamatsu Photonics). The system was controlled by ImagePro5 Plus (Media Cybernetics) software which integrated time-lapse acquisition, stage movement, and software-based auto-focus. During the experiment, cells were grown and visualized in 12-well coverslip bottom plates (MatTek) coated with fibronectin (Sigma).

Image analysis of time-lapse movies. Custom image analysis software described in our previous study[24] performed cell tracking and segmentation, and background and bleaching corrections. Watershed segmentation was applied to images after flat-field correction and background subtraction. Tracking of cells was performed by analysing the movie from end to start and linking each segmented cell to the cell in the previous image with the closest centroid. Cell divisions were automatically detected by a sharp twofold drop in total fluorescence between consecutive images.

Auto-correlation calculation. The auto-correlation function was computed with a robust estimator $A(\tau) = \langle\langle X_i(t)\cdot X_i(t+\tau)\rangle_t\rangle_i / \langle\langle X_i(t)^2\rangle_t\rangle_i$, where averages over time and over cells are denoted $<>_t$ and $<>_i$, respectively. We measured the auto-correlation of the ranked total fluorescence $R_i(t)$, such that $X_i(t) = R_i(t) - \langle R_i(t)\rangle_i$. This method was less sensitive to outliers in the data than auto-correlation over the total fluorescence (see Supplementary Information).

Correlations of tagged protein pairs. *mCherry*[29] was introduced into the genome on a retroviral vector identical to the *YFP* vector, but with *mCherry* in reading frame 0 instead of *EYFP*. Cells with YFP-tagged proteins were super-infected with this vector and sorted for red fluorescence to produce double-labelled cell clones.

Red-tagged proteins were identified with 3′-RACE (for details see Supplementary Information). About 1,000 double-labelled clones were screened for ribosomal or nuclear localization patterns, to obtain the clones studied here. Nuclear fluorescence levels of both colours were obtained from snapshots of cells, defining the nucleus by means of Hoechst 33342 ($5\,ng\,ml^{-1}$) staining of DNA. Spill-over of emission from Hoechst 33342 into the YFP or mCherry channels and between tagged proteins in the same cell was negligible (<0.1% of total signal). 3′-RACE with both *YFP* and *mCherry* primers revealed that different alleles of RPL5 were labelled, with both integrations in intron-1. Normalized mCherry and YFP fluorescence distributions for RPL5 were not significantly different (Kolmogorov–Smirnov test, $P = 0.75$).

Received 8 May; accepted 2 October 2006.
Published online 19 November 2006.

1. Spudich, J. L. & Koshland, D. E. Jr. Non-genetic individuality: chance in the single cell. *Nature* **262,** 467–471 (1976).
2. Novick, A. & Weiner, M. Enzyme induction as an all-or-none phenomenon. *Proc. Natl Acad. Sci. USA* **43,** 553–566 (1957).
3. Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman & Hall/CRC Press, Boca Raton, Florida, 2006).
4. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94,** 814–819 (1997).
5. Ferrell, J. E. Jr & Machleder, E. M. The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes. *Science* **280,** 895–898 (1998).
6. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial persistence as a phenotypic switch. *Science* **305,** 1622–1625 (2004).
7. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297,** 1183–1186 (2002).
8. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nature Genet.* **38,** 636–643 (2006).
9. Newman, J. R. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441,** 840–846 (2006).
10. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31,** 69–73 (2002).
11. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307,** 1965–1969 (2005).
12. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307,** 1962–1965 (2005).
13. Austin, D. W. *et al.* Gene network shaping of inherent noise spectra. *Nature* **439,** 608–611 (2006).
14. Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304,** 1811–1814 (2004).
15. Mihalcescu, I., Hsing, W. & Leibler, S. Resilient circadian oscillator revealed in individual cyanobacteria. *Nature* **430,** 81–85 (2004).
16. Colman-Lerner, A. *et al.* Regulated cell-to-cell variation in a cell-fate decision system. *Nature* **437,** 699–706 (2005).
17. Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J. & Eisen, M. B. Noise minimization in eukaryotic gene expression. *PLoS Biol.* **2,** e137 (2004).
18. Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422,** 633–637 (2003).
19. Acar, M., Becskei, A. & van Oudenaarden, A. Enhancement of cellular memory by reducing stochastic transitions. *Nature* **435,** 228–232 (2005).
20. Volfson, D. *et al.* Origins of extrinsic variability in eukaryotic gene expression. *Nature* **439,** 861–864 (2005).
21. Hooshangi, S., Thiberge, S. & Weiss, R. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc. Natl Acad. Sci. USA* **102,** 3581–3586 (2005).
22. Kærn, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Rev. Genet.* **6,** 451–464 (2005).
23. Paulsson, J. Summing up the noise in gene networks. *Nature* **427,** 415–418 (2004).
24. Sigal, A. *et al.* Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. *Nature Methods* **3,** 525–531 (2006).
25. Jarvik, J. W. *et al. In vivo* functional proteomics: mammalian genome annotation using CD-tagging. *Biotechniques* **33,** 852–4, 856 858–60 (2002).
26. Morin, X., Daneman, R., Zavortink, M. & Chia, W. A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in *Drosophila. Proc. Natl Acad. Sci. USA* **98,** 15050–15055 (2001).
27. Clyne, P. J., Brotman, J. S., Sweeney, S. T. & Davis, G. Green fluorescent protein tagging *Drosophila* proteins at their native genomic loci with small *P* elements. *Genetics* **165,** 1433–1441 (2003).
28. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123,** 1025–1036 (2005).
29. Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature Biotechnol.* **22,** 1567–1572 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Kahn Family Foundation and the Israel Science Foundation for support. We thank J. Paulsson, J. Pedraza and A. Eldar for discussions of the manuscript, and A. Sharp and E. Ariel for flow cytometry assistance. R.M. thanks the Horowitz Complexity Science Foundation for support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to U.A. (uri.alon@weizmann.ac.il).