# Central Dogma at the Single-Molecule Level in Living Cells

# Central dogma at the single-molecule level in living cells

Gene-Wei Li[1]† & X. Sunney Xie[1]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, USA 02138. †Present address: Department of Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, California, USA 94158.

**Gene expression originates from individual DNA molecules within living cells. Like many single-molecule processes, gene expression and regulation are stochastic, that is, sporadic in time. This leads to heterogeneity in the messenger RNA and protein copy numbers in a population of cells with identical genomes. With advanced single-cell fluorescence microscopy, it is now possible to quantify transcriptomes and proteomes with single-molecule sensitivity. Dynamic processes such as transcription factor binding, transcription and translation can be monitored in real time, providing quantitative descriptions of gene expression and regulation, and the demonstration that a single-molecule event can determine the phenotype of a cell.**

This year marks the thirty-fifth anniversary of single-molecule optical detection and imaging. In 1976, Thomas Hirschfeld successfully detected single molecules at room temperature using an optical microscope to reduce probe volume and hence the background signal[1]. Figure 1a shows his one-dimensional (1D) fluorescence image of individual immobilized protein molecules, each labelled with tens of fluorophores. The use of tightly focused laser beams eventually allowed single-fluorophore detection in solution phase at room temperature in 1990[2]. Imaging of single fluorophores in ambient environments was first reported with a scanning probe method in 1993[3], followed by much easier and improved methods[4-8] akin to Hirschfeld's that remain the methods of choice for imaging single molecules until the present. In the past decade, improvements in photodetectors and optical components have enabled extensive single-molecule fluorescence studies on a variety of biological problems first *in vitro* and more recently in living cells.

In a single-molecule experiment, one often observes stochastic behaviour, which would be otherwise obscured in an ensemble measurement. Figure 1b shows an early real-time observation of enzymatic turnovers of a single enzyme molecule, cholesterol oxidase[9]. The enzyme contains a flavin moiety that is naturally fluorescent in its oxidized form, but not fluorescent in its reduced form. Each on/off cycle of fluorescence emission corresponds to an enzymatic turnover. This time trace resembles the electrical signal of a single ion channel recorded using a patch clamp — the first single-molecule technique in biology[10]. However, in this case, stochastic chemical reaction events of a single enzyme molecule are seen. Here, stochastic means that each

fluorescence on/off time is probabilistic. Unlike the deterministic chemical kinetics of ensembles, each time trace is different, although their statistical properties are reproducible. On a single-molecule basis, when a chemical reaction occurs, the formation of a chemical bond completes in less than 1 ps, the event of which cannot be resolved in a single-molecule experiment. However, the waiting time for the event is much longer and probabilistic. When the kinetic scheme of a reaction includes a rate-limiting step, the distribution of the waiting times follows a single exponential, and the number of events in a fixed time interval follows a Poisson distribution.

By contrast, if the overall reaction does not have one rate-limiting step but instead consists of identical sequential steps, the total waiting time is less stochastic. An example of this is DNA replication by a single DNA polymerase, which is the basis of single-molecule sequencing[11] and a key application of single-molecule enzymology in biotechnology. A stochastic time trace of individual nucleotides incorporated into a single-stranded DNA template by a single DNA polymerase molecule is shown in Fig. 1c. Although the waiting time for each base incorporation step is exponentially distributed, the total waiting time for replicating the long DNA is narrowly distributed[12], a consequence of the central limit theorem. Bacterial cell-cycle time, when limited by chromosome replication, is not stochastic for this reason[13]. The experiments in Fig. 1b and c were conducted under non-equilibrium steady-state conditions, in which the substrate concentration (thermodynamic driving force) does not change while substrate molecules are continuously converted to product molecules. This is similar to many non-equilibrium processes in a live cell, such as gene expression.

The central dogma of molecular biology states that genetic information encoded in DNA is transcribed to mRNA by RNA polymerase, and mRNA is translated to protein by ribosomes. In a living cell, DNA exists as individual molecules from which gene expression regulation originates. But our knowledge of gene expression has come mainly from genetic and biochemical studies conducted with large populations of cells and purified biomolecules, which often obscures the single-molecule nature of gene expression. In recent years, single-molecule experiments *in vitro* have provided mechanistic insight into the functions of macromolecules involved in gene expression, including transcriptional and translational machineries[14–18]. Compelling areas of further investigation involve the observation and quantitative description of gene expression and regulation in a living cell.

Not only is there only one copy (or a few copies) of a particular gene, but the copy number of a particular mRNA is also small owing to short intracellular mRNA lifetimes[19], at least in a bacterial cell. By contrast, the copy number of particular proteins range from zero to 10,000 (refs 20, 21); many important proteins, such as

transcription factors that regulate gene expression, have small copy numbers[22]. Consequently, single-molecule sensitivity of mRNA and protein is needed to quantify gene expression in individual cells.

Because of the stochasticity associated with the single or low copy-number macromolecules, gene expression of individual cells cannot be synchronized. It is therefore necessary to make real-time observations of gene expression and regulation in single living cells. In particular, stochastic binding or unbinding of transcription factors on a particular gene, when rate limiting, must result in stochastic mRNA production, just like the single enzyme traces in Fig. 1b, c. Stochastic degradation of individual mRNA molecules further contributes to fluctuations in protein production. These temporal fluctuations of the mRNA and protein numbers (see sketches in Fig. 2) result in cell-to-cell variation of the copy numbers, or gene expression 'noise'. Under the steady-state condition, the connection between temporal fluctuations and variation within the population is similar to ergodicity in statistical physics — the time average of a system equals the ensemble average of identical systems.

Here we review recent single-molecule experiments that provide quantitative descriptions of the central dogma in living bacterial cells, although the strategies and technical advances highlighted are applicable to future studies in eukaryotic cells. We show that single-molecule stochastic events have important biological consequences, such as determining the phenotype of a cell.

## Imaging single molecules in living cells

To image a particular biomolecule in a live cell with fluorescence microscopy, specific labelling is required. The advent of genetically encodable fluorescent proteins has provided the highest specificity so far, with minimal perturbation for live-cell imaging[23], allowing real-time observations of fusion proteins of interest. Although the weak signal of a single-fluorescent-protein molecule is detectable *in vitro* using a fluorescent microscope together with a combination of laser excitation and modern charge-coupled-device detectors (Fig. 3a), the use of single fluorescent-protein reporters in live cells is challenging due to strong cellular autofluorescence. This obstacle can be partly overcome by selecting fluorescent proteins that are spectrally separate from the autofluorescence, which is generally blue-green[24]. Yellow- or red-emitting fluorescent proteins are therefore favourable for live-cell single-molecule imaging. Furthermore, in the same spirit as Hirschfeld's experiment, signal can be improved by reducing the detection volume to minimize autofluorescence background,. For example, total internal reflection fluorescence microscopy (TIRFM) can limit the axial depth by illuminating with an evanescent wave that penetrates only a few hundred nanometres into a sample (Fig. 3b). TIRFM is therefore ideal for studying membrane protein dynamics[25,26], but it does not

allow imaging of the whole cell body. However, single-fluorescent-protein imaging using wide-field illumination is possible in bacterial cells, and their compact cell bodies make them ideal for single-molecule studies *in vivo* (Fig. 3a).

In living eukaryotic cells, imaging a single fluorescent protein is more difficult. A typical mammalian nucleus is 5–10 μm in diameter, compared with 1 μm for a bacterial cell. In a wide-field microscope, such a large cell volume gives rise to a strong out-of-focus autofluorescence signal, which overwhelms the signal of a single fluorescent protein. Probing DNA–protein interactions therefore requires 3D sectioning. Although confocal fluorescence microscopy with one-photon excitation could be used, it also causes photobleaching outside the focal plane[27]. One solution to this problem is to use two-photon fluorescence microscopy[28,29] (Fig. 3c), which allows localized excitation only at the laser focus, considerably reducing out-of-focus photobleaching while providing 3D sectioning in living eukaryotic cells. But, like confocal microscopy, it requires point scanning, thus limiting its time resolution. Alternatively, sheet illumination[30,31], in which a thin light sheet illuminates only the image plane (Fig. 3d), provides low fluorescence background and high sensitivity, as well as high temporal resolution, because it does not require point scanning. These techniques are being adapted for single-fluorescent-protein imaging in living eukaryotic cells.

In a bacterial cell, a freely diffusing protein is difficult to image because its fast diffusion spreads the signal throughout the whole cell[32,33]. However, if a single fluorescent protein is localized, it can be imaged above the cellular autofluorescence background[34]. This method, termed detection by localization (Fig. 4a), works as long as there is only one immobilized molecule in a diffraction limited volume (less than 10 molecules within a bacterial cell). Detection by localization can be done by tethering on membrane [34], specific or even transient non-specific binding on chromosome[35].

In cases where the frame rate of the camera is insufficient to detect transient localization (<10 ms), a shorter pulse of laser excitation can be provided with each imaging frame[33,35], an idea borrowed from strobe photography. Detection by localization therefore allows single-molecule observations with millisecond time resolution.

The width of a single molecule image is about half of the optical wavelength due to the diffraction limit. However the accuracy of determining the center position of a single isolated fluorescent protein can be as high as a few nanometer [34,36]. To image more concentrated samples, higher spatial precision can also be achieved by selectively observing only one molecule at a time using photo-activatable fluorescent proteins. This is the idea behind recent developments in single-molecule-based super-resolution imaging, such as stochastic optical reconstruction microscopy[37] and photoactivated localization microscopy[38,39], in which high-resolution images

are reconstructed from many single-molecule images. Future applications of super-resolution techniques will probably change the way we view intracellular processes[40] such as gene expression. Single-fluorophore detection as discussed above remains a prerequisite for super-resolution imaging.

## Transcription factor dynamics

As the first step of gene expression, transcription factors must bind to or unbind from DNA in response to environmental signals. Because transcription factors interact with DNA at one location, gene expression is stochastic when the binding and unbinding of a transcription factor becomes rate-limiting (Fig. 2). In the classic example of the *lac* operon, the transcription factor known as the *lac* repressor (LacI), which is expressed at less than five copies per cell[35], binds to or unbinds from operator sites to control transcription. With detection by localization, a single *lac* repressor fused to yellow fluorescent protein (YFP) can be visualized when bound to its operator in the *lac* operon[35]. When the inducer isopropyl-β-D-thiogalactoside (IPTG) is added to the cell, localized fluorescent spots disappear as a result of LacI dissociation (Fig. 4b). This live cell assay allows single-molecule measurements of transcription-factor dissociation kinetics.

In addition, the binding kinetics can be measured. When IPTG is removed from the medium, the localized signal reappears, indicating the rebinding of LacI (Fig. 4c). This experiment allowed the first measurement of the time required for a LacI molecule to find a vacant operator site on DNA. It takes less than 360 s for one repressor to search for one specific binding site[35].

The search time of 360 s is a result of complex molecular processes. The protein-DNA search problem was extensively studied in the 1970s and 1980s[41,42]. It was observed that the DNA-binding rate constant of transcription factors significantly exceeds that expected from the 3D diffusion limit for bimolecular binding[42,43]. This observation led to the prevailing model of facilitated diffusion. For a transcription factor or any DNA-binding protein to find a target sequence on DNA, it first binds somewhere along the DNA non-specifically and undergoes 1D diffusion in search of the target. If the target is not found, the transcription factor dissociates from the DNA to avoid long search time imposed by 1D diffusion . The 3D diffusion through cytoplasm is much faster, allowing the transcription factor to reach other far away segments of DNA quickly. This combined 1D and 3D search is repeated until the transcription factor finds the DNA segment containing the target sequence. With single-molecule experiments, one can probe these phenomena in real time and quantify the process.

In a series of single-molecule studies *in vitro*, 1D diffusion has been directly observed for fluorescently labelled transcription factors and other DNA-binding proteins along non-specific DNA under a microscope[44–49].

The observed 1D diffusion rate (on the order of 0.05 $\mu m^2 s^{-1}$) is much slower than the 3D diffusion in a live cell (~3 $\mu m^2 s^{-1}$) because the 1D diffusion of the transcription factor is coupled to simultaneous rotation around the DNA, such that the transcription factor tracks the pitch of the DNA double helix[50,51]. In the in vitro experiments, low salt concentrations were used to assure long non-specific residence time in order to record long trajectories of 1D diffusion. In a living cell, high salt concentration shortens the residence time, but the diffusion constant often remains the same[46]. Consequently the number of bases inspected in each 1D segment is reduced.

A key question is whether such facilitated diffusion occurs in live cells. Recent single-molecule experiments suggest that it does. We found that, during the search process, a transcription factor spends 90% of its time on non-specific DNA, and the residence time of non-specific binding is less than 5 ms (ref. 35). Given the 1D diffusion constant *in vitro*, the protein inspects ~100 base pairs (bp), which implies a 100-fold acceleration of target search compared to the case with no 1D diffusion[52]. This observation is consistent with mounting evidence[53–57] that the length of DNA segment a transcription factor inspects is shorter than 1,000 bp , the value estimated from early *in vitro* experiments[43]. This 100-bp range indicates that for a 5 x $10^6$ bp genome, a transcription factor must inspect 5 x $10^6$/100 = 5 x $10^4$ segments before reaching the target site. Therefore, the total search time for one transcription factor in a cell is ~5 x $10^4$ x 5 ms = 250 s, in close agreement with the measured search time[35].

The combination of different single-molecule approaches have resolved the search problem and led to a quantitative understanding of the facilitated diffusion of transcription factors in bacteria. Similar single-molecule experiments should be able to address the same search problem in mammalian cells, which are complicated by nucleosomes[52].

## Translation in real time

Transcription-factor binding or unbinding leads to transcription and translation. Although the central dogma has been well established, real-time observation and quantitative description of transcription and translation in a single cell, at the single molecule level as motivated above, have only become possible in recent years. These studies have yielded unexpected observations of these fundamental processes in live cells[34,58–62].

We first discuss protein translation, as it can be relatively easily studied under repressed (uninduced) conditions, on a single-molecule basis. Under these conditions, single-molecule experiments have shown that proteins are synthesized in bursts[34], and that the characteristics of the bursts can be understood quantitatively at

the molecular level. The production of individual molecules of a YFP-fused membrane protein was monitored in real time in *E coli.*[34] (Fig. 4d). Newly synthesized YFPs were visualized one by one as diffraction-limited spots through detection by localization, and they were purposely photobleached after being detected. A fast-maturing YFP, Venus[63], was used to achieve seven-minute time resolution in the observation of translation. Using this approach, translational bursting from the *lac* operon under repressed conditions was observed[34]. Each burst creates four proteins on average, at a frequency of about one burst per generation time (although not synchronized to the cell cycle). The number of bursts per cell cycle follows the Poisson distribution -- a consequence of the exponentially distributed waiting times between each burst.

Because it was shown that each burst results from transcription of a single mRNA (generated due to the occasional dissociation of the LacI repressor), the observed translational burst must therefore be due to several rounds of ribosomal initiation on the same transcript. This transcript is degraded by nucleases with a stochastic cellular lifetime that is exponentially distributed with a time constant of 1.5 min. The longer an mRNA lives, the more proteins it produces. Consequently, as theoretically predicted in the 1970s[64–66], the burst size is exponentially distributed. This observation of exponentially distributed protein copy numbers per burst was independently confirmed by another single-molecule assay using β-galactosidase activity as a reporter[62]. As we discuss later, such stochastic expression due to transcription factor unbinding can be important in determining how a cell is induced in the presence of external stimuli[67].

## Transcription in real time

As discussed above, under repressed conditions, proteins are generated in bursts. At high expression levels, fluctuations of protein production in single cells have been extensively studied[68,69]. The origin of these fluctuations has been partly attributed to stochastic transcription and degradation of mRNA from the gene locus, but has not been directly visualized at the molecular level. Whereas the high copy number of proteins prevents the resolution of single copies, recent single-molecule studies of mRNA have shown that transcription is also burst-like[58–61]. But, in contrast to translational bursts, the origin of transcriptional bursts is largely unknown.

One widely used method to detect single mRNA molecules in living cells uses the bacteriophage coat proteins (MS2) that stably bind to specific RNA sequences[70]. To visualize single copies of mRNA, cell lines are engineered to express both MS2–GFP and mRNA containing several MS2-binding sites. First developed by the Singer group, this method allows real-time observation of transcript production, and is ideal for probing transcriptional dynamics in living cells by tracking and counting single mRNA–MS2–GFP complexes[71–72]. A

caveat is that the secondary structure associated with the binding sites and MS2 binding often interfere with the native mRNA degradation pathways[73], preventing the profiling of endogenous mRNA expression levels.

When MS2-containing mRNA is expressed under fully-induced conditions, the production of transcripts is found to be intermittent[58]. If transcript production were to have a single rate-limiting step, such as RNA polymerase binding or initiation, the waiting time between the birth of each mRNA would be exponentially distributed, and the copy-number distribution would be Poissonian (with a variance equal to the mean). Surprisingly, short bursts (average 6 min) of mRNA synthesis followed by long periods (average 37 min) of inactivity have been observed[58]. The burst-like transcription is similar to that shown in Fig. 2, even though there is no known transcription factor binding or unbinding in this case. This burst-like transcription was also observed using fluorescence correlation spectroscopy on MS2-bound mRNA in *Escherichia coli*[59], as well as in eukaryotic cells[60,61]. Although the overall waiting time between each mRNA synthesis event is not exponentially distributed[58], the waiting times for transition between the active and the inactive states are. Accordingly, the copy-number distribution is super-Poissonian, meaning that the variance of the distribution is greater than the mean. In other words, the cell-to-cell variation is significantly greater than would be expected from a single, rate-limiting process.

This important finding pointed out that transcription from a supposedly constitutive promoter is not as simple as RNA polymerases transcribing with a constant flux. Rather, it is a much noisier process, and the origin of this noise is unknown. Possible candidates include nucleoid-structuring proteins that are analogous to eukaryotic chromatin, global fluctuations of chromosome supercoiling states and RNA polymerase availability. *In vivo* single-molecule approaches are poised to further reveal the workings of these fundamental processes.

## Single-cell trajectories related to cell-to-cell variation

Under steady-state conditions, temporal fluctuations of gene expression in each cell lineage, as discussed in the previous sections and Fig. 2, lead to variation in copy number in an isogenic population of cells. A typical copy number distribution, which is often asymmetrical, is shown in Fig. 4e. A rigorous mathematical relationship between fluctuations in expression and the distribution of protein copy number in a population of cells was needed. A lognormal function has often been used as a convenient phenomenological fit, but it offers no physical insight.

For each gene, the dynamics of the central dogma can be described by two parameters — the burst frequency, $a$, which is the number of bursts per cell cycle, and the burst size, $b$, which is the average number of

molecules produced per burst. Experimentally, *a* and *b* can be determined by single-cell trajectories, such as in Fig. 2. Alternatively, the fact that temporal fluctuations in a cell lineage are related to cell-to-cell variation of copy numbers implies that *a* and *b* can also be inferred from for a population of isogenic cells at a particular moment as observed with a microscope or flow cytometer.

To find the relationship, one seeks a governing equation for gene-expression dynamics — the chemical master equation, which was first used by Delbruck[74] in 1940. In the late 1970s, the chemical master equation was applied to obtain protein copy-number distributions resulting from stochastic gene expression[64,65]. It was not until a decade ago that this approach regained attention [66,75,76], prompted by new experimental capabilities. Given the chemical kinetics scheme and rate constants connecting all the macromolecules involved in the central dogma, one can, in principle, solve the chemical master equation, which naturally yields time-dependent fluctuations. In practice, this can be simulated numerically using the Gillespie algorithm[77]. Under certain conditions, analytical results can be obtained. For example, under steady-state conditions with uncorrelated and exponentially distributed bursts, the chemical master equation can be solved[78], and the copy-number distribution, $p(n)$, can be approximated as gamma distribution when the copy number is approximated as a continuous variable[79]:

$$p(n) = n^{a-1} e^{-n/b} / b^a \Gamma(a)$$

The gamma distribution has two kinetic parameters — *a* and *b*, as defined earlier — providing a clear physical interpretation of the copy-number distribution.

This mathematical relationship allows extraction of intrinsic kinetics parameters (*a* and *b*) from fitting a gamma function to the measured copy-number distribution. At low expression levels, the values for *a* and *b* determined this way are consistent with those derived from the single-cell trajectories[34,62]. As shown in the next section, the cell-to-cell variation at high expression levels is more complicated but remains well described by a gamma distribution.

## Global profiling of expression variation

The ability to image single molecules in bacteria has offered an opportunity to profile protein expression globally, at any abundance levels. Pioneering work using a yeast green fluorescent protein (GFP) fusion library[80] surveyed the cell-to-cell variation of more than 2,500 genes under various growth conditions, yielding several important observations[81,82]. First, the noise, or the variance divided by the mean squared, scales inversely with abundance. Second, the deviation of noise in a particular gene away from the global trend

reflects its protein's function and perhaps the underlying regulation. However, because single-molecule sensitivity in yeast cells had not been achieved at the time, 30% of the genes that were weakly expressed in the GFP library were left undetectable.

To profile global variation at all expression levels, an *E. coli* YFP fusion library was constructed, which included more than 1,000 genes with expression levels ranging from 0.1 to $10^4$ per cell[21]. Of all the tagged proteins, approximately 99% of the copy-number distributions are well fit with the gamma distribution. 50% of the proteins are expressed at an average level of less than ten molecules per cell, which argues for the necessity of single molecule sensitivity in single cell analyses. Protein-expression noise has two distinct scaling properties relative to the mean. Below ten molecules per cell, the noise is inversely proportional to protein abundance. This scaling is the same as what was observed in yeast, indicating that the noise from random birth and death of molecules, also known as intrinsic noise[83,84], dominates the expression variation for low-abundance proteins. By contrast, at abundances above ten molecules per cell, the noise reaches a plateau of 30% and does not decrease any further. This noise plateau is common, or 'extrinsic' to most high-abundance proteins, as the expression levels of different proteins have a large covariance from cell-to-cell. Notably, time-lapse movies have shown that the extrinsic noise fluctuates at a timescale much longer than the cell cycle, suggesting that a slow global-regulation process is at work[68].

At the transcriptional level, the same YFP library was used to simultaneously survey mRNA and protein variation for 137 genes[21]. Instead of labelling with MS2, which requires further cloning steps, mRNA was visualized using single-molecule fluorescence *in situ* hybridization (FISH)[85] in fixed cells. Unlike conventional approaches that use several hybridization probes against the mRNA, the YFP mRNA was targeted using a universal singly labelled FISH probe optimized for both hybridization efficiency towards its targets and specificity against off-targets. It was found that, even for highly expressed genes, the average mRNA copy number is less than five per cell. Among a population of genetically identical cells, every mRNA species has a super-Poissonian distribution, which is related to the transcriptional bursts observed in the real-time experiments and suggests that this is a general phenomenon for most genes.

The simultaneous profiling of mRNA and protein[21] also showed that the mRNA and protein copy numbers of a single cell for any given gene are uncorrelated; that is, a cell that has more mRNA molecules than average does not necessarily have more proteins (Fig. 4e). This perhaps counterintuitive result can be explained by the fact that mRNA has a much shorter lifetime than protein in bacteria[19]. This finding argues for the necessity of single-cell proteomics analyses, and offers a warning for interpretations of single-cell transcriptome

analyses, at least for bacteria. A mammalian cell, on the other hand, has comparable mRNA and protein lifetimes, and hence is expected to have more correlated mRNA and protein levels than in a bacterial cell.

## Single-molecule event triggers phenotypic switching

How cells with identical genomes have different phenotypes is an interesting question. Phenotypes are the physical, chemical and physiological states of the cell as related to function, determined by both the genome and environment. Given the ubiquitous and substantial noise described earlier, it is evident that the phenotype of a cell cannot be solely defined by its transcriptome and proteome. Cells can tolerate rather large noise of protein and mRNA abundance while tightly maintaining their phenotypes. A compelling question is what molecular actions dictate the transition between phenotypes.

In some cases, the cell phenotype can be clearly defined when there are bimodal or multimodal distributions of proteins, in contrast to the unimodal copy-number distribution that are most often observed[21]. As shown in Fig. 5b, a population of isogenic *E. coli* cells, in which *lac* permease is labelled with YFP, shows bistability. The *lac* operon in *E. coli*, consisting of *lacZ*, *lacY* and *lacA* genes, is normally repressed by the *lac* transcription-factor repressor (LacI) in the absence of inducer (Fig. 5a). When the inducer is present, it inactivates LacI and triggers expression of the *lac* operon. The synthesis of the permease increases the inducer influx that inactivates more LacI repressor, creating a positive feedback on permease expression[86]. Without inducer, no cells are induced, whereas with high inducer concentrations, all cells are induced. At moderate inducer concentrations, only a fraction of the cells is induced. This bistability is controlled by the positive feedback of the *lac* operon.

Bistability is commonly exploited by bacteria to generate alternative phenotypes[87], such as persistence against antibiotics[88], lysis or lysogeny after phage infection[89] and induction of the *lac* operon in *E. coli*[86]. Although much is known about the genetic switches, what drives the transition between two phenotypic states is unclear in many cases. How does a single cell make a decision on which phenotype to choose? With single-molecule imaging, uninduced *E. coli* cells have been shown to contain 0–10 copies of the permease enzyme, which is below the threshold for positive feedback (more than 300 molecules per cell)[67]. Transition to the fully induced state therefore requires a large burst of protein synthesis (Fig. 5c).

The transcription factor controlling permease synthesis, the LacI repressor, is a tetramer that binds to two DNA-binding sites, creating a DNA loop. Partial dissociation of LacI and rapid rebinding to DNA result in a single copy of mRNA and a small burst of permease, as was observed in the aforementioned real-time studies of

the repressed *lac* promoters. When the repressor completely dissociates from both operators on DNA, a large burst of permease arises, because it takes a few minutes for the repressor to rebind[35]. Indeed, bistability was eliminated in strains without DNA looping[67]. It is the stochastic single-molecule event of complete repressor dissociation from DNA that triggers the cell's phenotypic switching.

## Looking forward

We have shown that in the case of *lac* operon, the workings of the genetic switch can be quantitatively understood at the molecular level. This is an example that low probability, stochastic events of a single molecule have important biological consequences. Another trivial example is point mutations in the course of evolution..

It is well recognized that such stochastic events are connected to cell-fate determination of other systems[90] For example, there is considerable evidence that bacterial persistence against antibiotics is a stochastic process involving gene expression[88]. Persisters are not drug resistant but are drug tolerant. Drug resistance is related to a changing genome, whereas persisters have identical genomes, but different phenotypes. The phenomenon exists for many bacterial species and antibiotics. The molecular mechanism behind persistence is largely unknown, partly because the tools are not available. Understanding the molecular mechanism of persistence may be crucial to drug development, especially for diseases such as tuberculosis, caused by the bacterium *Mycobacterium tuberculosis*, which kills almost two million people every year worldwide. Single-cell gene-expression profiling may shed light on the mechanism of persistence.

Similarly, the reprogramming of somatic cells into induced pluripotent stem cells in the presence of certain transcription factors is also stochastic[91]. There are no elite cells and every cell has a certain probability of being reprogrammed,in the presence of some transcription factors, which is analogous to stochastic switching in *E. coli* lac operon at low inducer concentrations. Yet, unlike the *lac* operon, the molecular mechanism is unknown. Extension of single-molecule approaches into mammalian cells and stem cells will allow real-time monitoring over long periods so that low probability events with considerable biological consequences can be observed directly. We anticipate that the single-molecule approaches summarized in this Review will lead to more biological discoveries for many years to come.

**Figure 1 | Stochastic nature of single-molecule processes. a**, Optical imaging of single protein molecules at room temperature. In his 1976 work, Hirschfeld demonstrated the detection of single protein molecules using a fluorescence microscope. A line scan of eight protein molecules was recorded. Adapted from ref. 1. **b**, Stochastic turnovers of a single enzyme molecule. The fluorescence signal of a cholesterol oxidase molecule exhibits stochastic switching between a fluorescent (reduced flavin) and nonfluorescent (oxidized flavin) state as enzymatic turnovers take place. The waiting time before each switching event is highly variable owing to a single rate-limiting step. Adapted from ref. 9. **c**, Single-molecule DNA sequencing. A single DNA polymerase is used to sequence DNA by incorporating fluorescently labelled nucleotides of four different colors. Although each incorporation happens stochastically with variable waiting times, the overall time for DNA replication, which is a sum of many sequential steps, is narrowly distributed. a.u., arbitrary units; adapted from ref. 11.

**Figure 2 | Central dogma at the single-molecule level.** In a living bacterial cell, there is usually one copy of a particular gene, which is regulated by transcription factors (TFs), and transcribed into mRNA and translated into protein. A rate-limiting event, such as TF binding and unbinding on DNA, in this single molecule process results in stochasticity. The expression levels of mRNA (middle panel) and protein (bottom panel) show temporal fluctuations in a single cell lineage. This gives rise to variations of mRNA and protein copy numbers among a population of cells at a particular time (right panels).

**Figure 3 | Methods for imaging single molecules in live cells.** Single-molecule fluorescence can be imaged using multiple laser illumination geometries that reduce the probe volume. **a**, In wide-field illumination, the entire cell is subject to laser exposure. For bacterial cells that have small volume, no further probe volume reduction is necessary. **b**, With total internal reflection, only the region within a few hundred nanometres from the coverslip is illuminated. This method is often used to image single membrane proteins, but cannot detect molecules deep in the cells. **c**, Two-photon excitation suppresses out-of-focus background, but suffers from slower time resolution owing to the need of point scanning **d**, Sheet illumination has reduced background, as well as increased time resolution because it does not require point scanning.

**Figure 4 | Real-time measurements of gene expression with single-molecule sensitivity. a**, Detection by localization. The cellular autofluorescence makes it difficult to detect a freely diffusing fluorescent protein. However, a localized single molecule can be imaged above the autofluorescence background. **b**, Detection of single transcription factors in live cells. A *lac* repressor (LacI) labelled with YFP can be imaged when bound to

its operator site on DNA. The localized fluorescence disappears after dissociation caused by the inducer IPTG. DIC, differential interference contrast microscopy; adapted from ref. 35. **c**, Target search by a transcription factor. When the inducer IPTG is removed (dilution), the *lac* repressor begins to search for its operator site. After rebinding to the target, localized fluorescence appears, allowing measurement of the search time. Adapted from ref. 35. **d**, Real time observation of protein synthesis at low expression levels. Individual YFP fusion protein molecules are visualized after being immobilized to the cell membrane, are synthesized in bursts after intrinsic noise. Adapted from ref. 34. **e**, Copy numbers of mRNA and protein of the same gene, measured in the same cell, show little correlation, which is mostly due to the differences in the lifetimes of mRNA and protein. The protein copy number distribution follows a gamma distribution. Adapted from ref. 21.

**Figure 5 | Phenotype switching due to a single-molecule event. a**, Bistability of the *lac* operon. The positive feedback by the normally repressed *lac* permease (LacY) results in bimodal distribution at intermediate inducer concentration with two distinct phenotypes, fluorescent or not. **b**, Threshold for positive feedback. The copy-number distribution of LacY in uninduced cells is the same as that without inducer, suggesting that the typical leaky expression of LacY is not sufficient to trigger the positive feedback. **c**, A large burst of expression originates from the stochastic event of the complete dissociation of a single transcription factor from DNA. A large expression burst of LacY (~300 molecules) is necessary to trigger the positive feedback. The switching of the phenotype is attributed to the complete dissociation of a single transcription factor, LacI. This experiment shows that a low probability single-molecule event can determine cell fate. Adapted from ref. 67.

1. Hirschfeld, T. Optical microscopic observation of single small molecules. *Appl. Opt.* **15,** 2965–2966 (1976).

**This paper presents the first single-molecule fluorescence imaging at room temperature, and the approach — reducing the detection volume to enhance signal-to-background ratio — has helped lead to today's single-molecule studies in living cells.**

2. Shera, E. B., Seitzinger, N. K., Davis, L. M., Keller, R. A. & Soper, S. A. Detection of single fluorescent molecules. *Chem. Phys. Lett.* **174,** 553–557 (1990).

3. Betzig, E. & Chichester, R. J. Single molecules observed by near-field scanning optical microscopy. *Science* **262**, 1422-1425 (1993).

4. Funatsu, T., Harada, Y., Tokunaga, M., Saito, K. & Yanagida, T. Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution. *Nature* **374,** 555–559 (1995).

5. Schmidt, T., Schutz, G. J., Baumgartner, W., Gruber, H. J. & Schindler, H. Characterization of photophysics and mobility of single molecules in a fluid lipid-membrane. *J. Phys. Chem.* **99,** 17662–17668 (1995).

6. Xie, X. S. & Trautman, J. K. Optical studies of single molecules at room temperature. *Annu. Rev. Phys. Chem.* **49,** 441–480 (1998).

7. Moerner, W. E. & Orrit, M. Illuminating single molecules in condensed matter. *Science* **283,** 1670–1676 (1999).

8. Ha, T. *et al.* Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proc. Natl. Acad. Sci. USA* **96,** 893–898 (1999).

9. Lu, H. P., Xun, L. & Xie, X. S. Single-molecule enzymatic dynamics. *Science* **282,** 1877–1882 (1998).

10. Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflugers Arch.* **391,** 85–100 (1981).

11. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

12. Bouche, J. P. Physical map of a 470 x $10^3$ base-pair region flanking the terminus of DNA replication in the *Escherichia coli* K12 genome. *J. Mol. Biol.* **154,** 1–20 (1982).

13. Bremer, H. A stochastic process determines the time at which cell division begins in Escherichia coli. *J. Theor. Biol.* **118,** 351–365 (1986).

14. Wang, M. D. *et al.* Force and velocity measured for single molecules of RNA polymerase. *Science* **282,** 902–907 (1998).

15. Blanchard, S. C., Kim, H. D., Gonzalez, R. L., Jr., Puglisi, J. D. & Chu, S. tRNA dynamics on the ribosome during translation. *Proc. Natl Acad. Sci. USA* **101,** 12893–12898 (2004).

16. Kapanidis, A. N. *et al.* Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science* **314,** 1144–1147 (2006).

17. Wen, J. D. *et al.* Following translation by single ribosomes one codon at a time. *Nature* **452,** 598–603 (2008).

18. Cornish, P. V., Ermolenko, D. N., Noller, H. F. & Ha, T. Spontaneous intersubunit rotation in single ribosomes. *Mol. Cell* **30,** 578–588 (2008).

19. Bernstein, J. A., Khodursky, A. B., Lin, P. H., Lin-Chao, S. & Cohen, S. N. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl Acad. Sci. USA* **99,** 9697–9702 (2002).

20. Ishihama, Y. *et al.* Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* **9,** 102 (2008).

21. Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329,** 533–538 (2010).

**This article profiles the transcriptome and proteome with single-molecule sensitivity in single cells, and shows that the protein copy-number distribution of most genes can be well fit with gamma distribution.**

22. Xie, X. S., Choi, P. J., Li, G. W., Lee, N. K. & Lia, G. Single-molecule approach to molecular biology in living bacterial cells. *Annu. Rev. Biophys.* **37,** 417–444 (2008).

23. Giepmans, B. N., Adams, S. R., Ellisman, M. H. & Tsien, R. Y. The fluorescent toolbox for assessing protein location and function. *Science* **312,** 217–224 (2006).

24. Andersson, H., Baechi, T., Hoechl, M. & Richter, C. Autofluorescence of living cells. *J. Microsc.* **191,** 1–7 (1998).

25. Sako, Y., Minoghchi, S. & Yanagida, T. Single-molecule imaging of EGFR signalling on the surface of living cells. *Nature Cell Biol.* **2,** 168–172 (2000).

26. Iino, R., Koyama, I. & Kusumi, A. Single molecule imaging of green fluorescent proteins in living cells: E-cadherin forms oligomers on the free cell surface. *Biophys. J.* **80,** 2667–2677 (2001).

27. Wilson, T. *Confocal Microscopy* (Academic Press, 1990).

28. Denk, W., Strickler, J. H. & Webb, W. W. Two-photon laser scanning fluorescence microscopy. *Science* **248,** 73–76 (1990).

29. Sanchez, E. J., Novotny, L., Holtom, G. R. & Xie, X. S. Room-temperature fluorescence imaging and spectroscopy of single molecules by two-photon excitation. *J Phys Chem A* **101**, 7019-7023 (1997).

30. Fuchs, E., Jaffe, J., Long, R. & Azam, F. Thin laser light sheet microscope for microbial oceanography. *Opt. Express* **10,** 145–154 (2002).

31. Ritter, J. G., Veith, R., Veenendaal, A., Siebrasse, J. P. & Kubitscheck, U. Light sheet microscopy for single molecule tracking in living tissue. *PLoS ONE* **5,** e11639 (2010).

32. Elowitz, M. B., Surette, M. G., Wolf, P. E., Stock, J. B. & Leibler, S. Protein mobility in the cytoplasm of Escherichia coli. *J. Bacteriol.* **181,** 197–203 (1999).

33. Xie, X. S., Yu, J. & Yang, W. Y. Living cells as test tubes. *Science* **312,** 228–230 (2006).

34. Yu, J., Xiao, J., Ren, X., Lao, K. & Xie, X. S. Probing gene expression in live cells, one protein molecule at a time. *Science* **311,** 1600–1603 (2006).

**This article monitors the real-time production of individual proteins by imaging newly synthesized YFP molecules, revealing translational bursts from a repressed promoter.**

35. Elf, J., Li, G. W. & Xie, X. S. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316,** 1191–1194 (2007).

**This article probes gene-regulation dynamics by following the search time of transcription factors to their targets on DNA in living cells.**

36. Thompson, R. E., Larson, D. R. & Webb, W. W. Precise nanometer localization analysis for individual fluorescent probes. *Biophys. J.* **82,** 2775–2783 (2002).

37. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods* **3,** 793–795 (2006).

38. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313,** 1642–1645 (2006).

39. Hess, S. T., Girirajan, T. P. & Mason, M. D. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91,** 4258–4272 (2006).

40. Hell, S. W. Far-field optical nanoscopy. *Science* **316,** 1153–1158 (2007).

41. Richter, P. H. & Eigen, M. Diffusion controlled reaction rates in spheroidal geometry. Application to repressor—operator association and membrane bound enzymes. *Biophys. Chem.* **2,** 255–263 (1974).

42. Berg, O. G., Winter, R. B. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **20,** 6929–6948 (1981).

43. Riggs, A. D., Bourgeois, S. & Cohn, M. The lac repressor-operator interaction. 3. Kinetic studies. *J. Mol. Biol.* **53,** 401–417 (1970).

44. Kabata, H. *et al.* Visualization of single molecules of RNA polymerase sliding along DNA. *Science* **262,** 1561–1563 (1993).

45. Harada, Y. *et al.* Single-molecule imaging of RNA polymerase-DNA interactions in real time. *Biophys. J.* **76,** 709–715 (1999).

46. Blainey, P. C., van Oijen, A. M., Banerjee, A., Verdine, G. L. & Xie, X. S. A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc. Natl Acad. Sci. USA* **103,** 5752–5757 (2006).

47. Graneli, A., Yeykal, C. C., Robertson, R. B. & Greene, E. C. Long-distance lateral diffusion of human Rad51 on double-stranded DNA. *Proc. Natl Acad. Sci. USA* **103,** 1221–1226 (2006).

48. Wang, Y. M., Austin, R. H. & Cox, E. C. Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys. Rev. Lett.* **97,** 048302 (2006).

49. Kim, J. H. & Larson, R. G. Single-molecule analysis of 1D diffusion and transcription elongation of T7 RNA polymerase along individual stretched DNA molecules. *Nucleic Acids Res.* **35,** 3848–3858 (2007).

50. Schurr, J. M. The one-dimensional diffusion coefficient of proteins absorbed on DNA. Hydrodynamic considerations. *Biophys. Chem.* **9,** 413–414 (1979).

51. Blainey, P. C. *et al.* Nonspecifically bound proteins spin while diffusing along DNA. *Nature Struct. Mol. Biol.* **16,** 1224–1229 (2009).

52. Li, G. W., Berg, O. G. & Elf, J. Effects of macromolecular crowding and DNA looping on gene regulation kinetics. *Nat Phys* **5**, 294-297 (2009).

53. Stanford, N. P., Szczelkun, M. D., Marko, J. F. & Halford, S. E. One- and three-dimensional pathways for proteins to reach specific DNA sites. *EMBO J.* **19,** 6546–6557 (2000).

54. Gowers, D. M. & Halford, S. E. Protein motion from non-specific to specific DNA by three-dimensional routes aided by supercoiling. *EMBO J.* **22,** 1410–1418 (2003).

55. Bonnet, I. *et al.* Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA. *Nucleic Acids Res.* **36,** 4118–4127 (2008).

56. Porecha, R. H. & Stivers, J. T. Uracil DNA glycosylase uses DNA hopping and short-range sliding to trap extrahelical uracils. *Proc. Natl Acad. Sci. USA* **105,** 10791–10796 (2008).

57. Winter, R. B., Berg, O. G. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli lac* repressor–operator interaction: kinetic measurements and conclusions. *Biochemistry* **20,** 6961–6977 (1981).

58. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123,** 1025–1036 (2005).

**This article uses MS2–GFP to follow the production of individual mRNA molecules in real time.**

59. Le, T. T. *et al.* Real-time RNA profiling within a single bacterium. *Proc. Natl Acad. Sci. USA* **102,** 9160–9164 (2005).

60. Chubb, J. R., Trcek, T., Shenoy, S. M. & Singer, R. H. Transcriptional pulsing of a developmental gene. *Curr. Biol.* **16,** 1018–1025 (2006).

61. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4,** e309 (2006).

62. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440,** 358–362 (2006).

63. Nagai, T. *et al.* A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature Biotechnol.* **20,** 87–90 (2002).

64. Berg, O. G. A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.* **71,** 587–603 (1978).

65. Rigney, D. R. Stochastic-model of constitutive protein-levels in growing and dividing bacterial-cells. *J. Theor. Biol.* **76,** 453–480 (1979).

66. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94,** 814–819 (1997).

67. Choi, P. J., Cai, L., Frieda, K. & Xie, X. S. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science* **322,** 442–446 (2008).

**This article investigates the triggering mechanism of the *lac* operon bistable switch and shows that the dissociation of a single *lac* repressor is responsible for the phenotype switch to the induced state.**

68. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297,** 1183–1186 (2002).

69. 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).

70. Tyagi, S. Imaging intracellular RNA distribution and dynamics in living cells. *Nature Methods* **6,** 331–338 (2009).

71. Bertrand, E. *et al.* Localization of *ASH1* mRNA particles in living yeast. *Mol. Cell* **2,** 437–445 (1998).

**This study develops an approach to image individual mRNA molecules in living cells, using fluorescent protein-fused MS2 proteins that bind to an engineered mRNA sequence.**

72. Beach, D. L., Salmon, E. D. & Bloom, K. Localization and anchoring of mRNA in budding yeast. *Curr. Biol.* **9,** 569–578 (1999).

73. Johansson, H. E. *et al.* A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein. *Proc. Natl Acad. Sci. USA* **95,** 9244–9249 (1998).

74. Delbrüück, M. Statistical fluctuations in autocatalytic reactions. *J. Chem. Phys.* **8,** 120–140 (1940).

75. Paulsson, J. Summing up the noise in gene networks. *Nature* **427,** 415–418 (2004).

76. Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proc. Natl Acad. Sci. USA* **105,** 17256–17261 (2008).

77. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.* **22,** 403–434 (1976).

78. Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys. Rev. Lett.* **84,** 5447–5450 (2000).

79. Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* **97,** 168302 (2006).

**This article derives gamma distribution of protein copy numbers in a model with burst-like gene expression; the distribution is in agreement with experimental observations in reference 62.**

80. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691 (2003).

81. Newman, J. R. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441,** 840–846 (2006).

82. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nature Genet.* **38,** 636–643 (2006).

83. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl Acad. Sci. USA* **99,** 12795–12800 (2002).

84. Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc. Natl Acad. Sci. USA* **98,** 8614–8619 (2001).

85. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts *in situ*. *Science* **280,** 585–590 (1998).

86. Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & Van Oudenaarden, A. Multistability in the lactose utilization network of *Escherichia coli*. *Nature* **427,** 737–740 (2004).

87. Dubnau, D. & Losick, R. Bistability in bacteria. *Mol. Microbiol.* **61,** 564-–572 (2006).

88. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial persistence as a phenotypic switch. *Science* **305,** 1622–1625 (2004).

89. Ptashne, M. *A Genetic Switch* 3 edn (Cold Spring Harbor Laboratory Press, 2004).

90. Losick, R. & Desplan, C. Stochasticity and cell fate. *Science* **320,** 65–68 (2008).

91. Hanna, J. *et al.* Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462,** 595–601 (2009).

**Acknowledgements** We thank current and former members of Xie group for their contributions to the work described herein, and W. Greenleaf, P. Sims, L. Cai, P. Choi, H. Ge and J. Elf for their critical comments on this review. This work is supported by the National Institutes of Health Director's Pioneer Award.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to X.S.X. (xie@chemistry.harvard.edu).