Stochastic gene expression: from single molecules to the proteome

Benjamin B Kaufmann[1,2] and Alexander van Oudenaarden[1]

Protein production involves a series of stochastic chemical steps. One consequence of this fact is that the copy number of any given protein varies substantially from cell to cell, even within isogenic populations. Recent experiments have measured this variation for thousands of different proteins, revealing a linear relationship between variance and mean level of expression for much of the proteome. This simple relationship is frequently thought to arise from the random production and degradation of mRNAs, but several lines of evidence suggest that infrequent gene activation events also bear responsibility. In support of the latter hypothesis, single-molecule experiments have demonstrated that mRNA transcripts are often produced in large bursts. Moreover, the temporal pattern of these bursts appears to be correlated for chromosomally proximal genes, suggesting the existence of an upstream player.

Addresses
[1] Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2] Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

Corresponding author: van Oudenaarden, Alexander (avo1@mit.edu)

Current Opinion in Genetics & Development 2007, **17**:107–112

This review comes from a themed issue on
Chromosomes and expression mechanisms
Edited by Tom Misteli and Abby Dernburg

Available online 20th February 2007

0959-437X/$ – see front matter

DOI 10.1016/j.gde.2007.02.007

Introduction

Within the confines of individual cells, minute changes in the concentration or spatial arrangement of molecular species can produce substantial effects. For example, a transcription factor equally prevalent in two *Escherichia coli* might be bound to a promoter in one and unbound in another, subject to the dictates of statistical mechanics. Protein production would consequently begin in one cell and not the other, amplifying the fluctuation and propelling each cell to a different fate. Identical genotype and growth environment are thus insufficient to ensure that two cells will develop the same phenotypes, an observation with a long history of investigation (for example, see [1–6]).

A major goal of recent research has been to identify and differentiate between the myriad possible origins of this variability, to understand which are important and which are not, and to put firm numbers on each of them. Although biochemical fluctuations influence all stages of gene expression, those involving molecules in extremely low abundance are expected from a statistical standpoint to be larger in magnitude and therefore to contribute disproportionately to the overall variation between cells. Validating this idea has required technological feats, described in part below, that include novel single-molecule assays and genome-wide noise surveys. As a result, a clear picture of which types of noise exist in cells and how they influence biological processes is now emerging.

Measuring noise

Noise experiments commonly begin with the insertion of a reporter gene (e.g. green fluorescent protein driven by a promoter of interest) into the genome. Cells are then cultured, usually in a swirling flask to provide a uniform environment. Finally, the fluorescence of many individual cells in that population is ascertained by microscopy or flow cytometry [7–12,13••,14••,15–17]. Alternatively, individual cells can also be followed over time, yielding important information on the dynamics of stochastic gene expression [10,16,18–24].

To quantify the heterogeneity of the population, the variance across the population divided by the mean squared is typically used, a parameter called the 'noise'. The specific measurable used to compute this differs from experiment to experiment, with options including total fluorescence, mean fluorescence, or fluorescence among cells that share similar morphological traits. Different choices tend to emphasize different aspects of cell-to-cell variation. Measurements using total cell fluorescence, for example, invariably include not only stochasticity in gene expression but also variation in cell size, and therefore best describe the diversity of a growing asynchronous population. Selecting cells all at one stage in the cell cycle, by contrast, reduces this 'growth' heterogeneity, enabling more sensitive detection of the underlying biochemical noise [11,12,14••,15,24,25,26••].

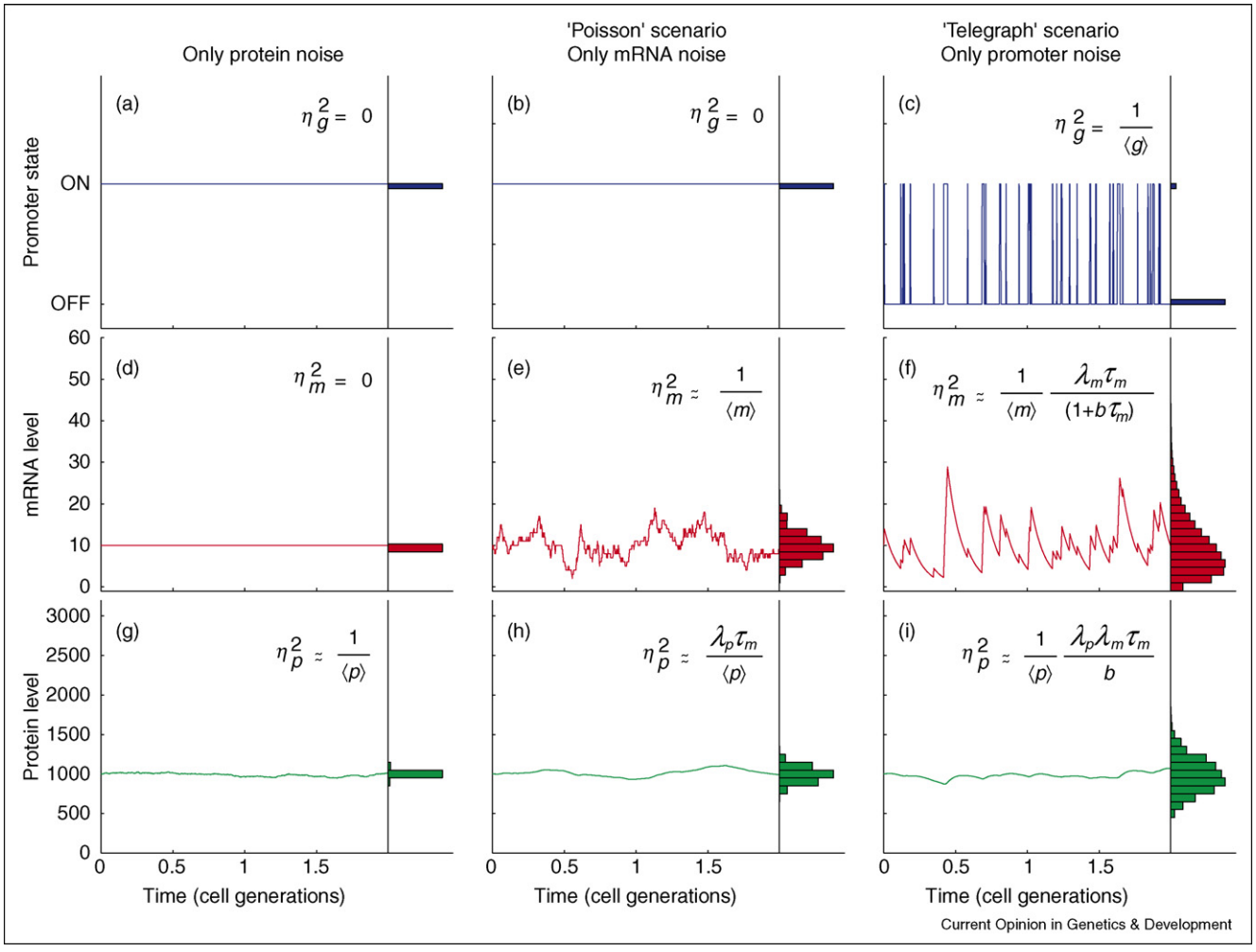
mRNA fluctuations

Individual studies have targeted many candidate noise sources, providing important proofs of principle that noise in gene expression is influenced by numerous parameters such as chromosomal position [8,14••,27•], the presence of upstream regulators [8–11], and 'global' factors that impact all genes or proteins *en masse* [10–12,13••,14••,16,25]. But which are most important? Two recent studies, both in

Saccharomyces cerevisiae, have broadened our perspective in this regard. Newman *et al.* [14••] used high-throughput flow cytometry to examine more than 2500 green fluorescent protein (GFP)-tagged fusion proteins, providing a broad survey of protein noise. Bar-Even *et al.* [13••] measured only 43 fusions, but each was examined in 11 different experimental conditions, and the authors combined their measurements with a comprehensive modeling approach. Both studies made the same central observation: for the great majority of proteins, noise level is inversely proportional to the mean level of expression.

These findings support the long-held hypothesis [3,5,28–32] that protein noise originates in the random production and decay of low-copy mRNAs. In this view, referred to in Figure 1 as the 'Poisson' scenario, transcription occurs with constant probability over time, with single uncorrelated mRNAs being randomly produced and randomly

Figure 1



Comparison of three possible noise sources using the Gillespie algorithm. Rows show the activation state of the promoter **(a,b,c** blue**)**, the mRNA transcript level **(d,e,f** red**)** and protein level **(g,h,i** green**)**. The first vertical column describes a situation in which only protein creation and destruction are noisy. This scenario is not compatible with most experiments because the calculated noise levels are far lower than the measured values. The second column, referred to in the main text as the 'Poisson' scenario, assumes constant promoter activation (b) and includes noise only in mRNA production and destruction (e). Protein production then follows deterministically (h). The defining features of this scenario are an mRNA noise exactly equal to the inverse of the mean, and a protein noise that scales inversely with the mean [28]. The third column, described in the text as the 'Telegraph' scenario, assumes that the promoter becomes active only for short bursts (c) during which transcripts are made. In this figure, mRNA and protein production are then allowed to follow deterministically. Unlike the 'Poisson' scenario, mRNA noise is not equal to one over the mean, but is merely proportional (f). However, like the 'Poisson' scenario the protein level can under certain circumstances also be inversely proportional to the mean (see main text and (i)). In the equations, η^2 is the level of noise, defined as the variance divided by the mean squared. The subscripts g, m and p represent the fraction of the genes that are active, the mRNA level and the protein level, respectively. Angled brackets represent population averages. λ_m and λ_p are the mRNA and protein production rates, respectively, whereas τ_m is the mRNA half-life. b is the off rate of the activated promoter in the 'Telegraph' scenario, assumed here to be a constant.

destroyed. The defining feature of this model is that the steady-state mRNA levels across a population will follow a Poisson distribution, where noise equals exactly one over the mean number of mRNA molecules with no proportionality constant required (Figure 1e) [28]. Deterministic protein production would then amplify the mRNA fluctuations, giving rise to a protein noise that scales inversely with the mean with a proportionality constant equal to the average number of proteins produced per mRNA (Figure 1h), consistent with the experimentally observed scaling.

For this simple scaling to be true, the average number of proteins made per mRNA must not vary too dramatically from gene to gene; for example, as a result of mRNA-specific post-translational modifications or variations in ribosome binding affinity. Otherwise, protein production will not simply amplify mRNA noise by a constant that represents the average number of proteins produced per mRNA (see next section and Figure 1h). However, Bar-Even *et al.* [13••] were able to calculate a value, ~1200 proteins per mRNA, that allows a good fit to their data. Given that there appears to be no fundamental reason why such translational protein bursts should be approximately constant across the genome, an important test of this model will be to use single-molecule techniques to directly measure the protein burst sizes for the mRNAs examined in this study and across the yeast proteome to see if they can indeed be approximated by a constant.

The trend fails for highly transcribed genes. In this subset of genes, noise levels settle to a constant value that is independent of the mean level [13••,14••]. The likely reason is that mRNAs for these proteins are so abundant that noise in their production is overshadowed by other noise sources such as differences between cells in the number or concentration of essential machinery like polymerases, ribosomes or ATP [10–12,13••,14••,16]. Indeed, using a 'two-color' approach (see [12]), Newman *et al.* [14••] showed that the contribution mRNA fluctuations make to the total noise continued to follow the simple proportionality found for most genes.

Both studies also help substantiate the indirect population-level arguments of Fraser *et al.* [33] that show that genes whose deletion is lethal typically have lower noise levels than dispensable genes with comparable means. As specific examples, inessential stress genes are indeed found to be noisy, whereas those coding for indispensable protein degradation machinery show little cell-to-cell variation [13••,14••]. Together, these three studies thus argue that noise is biologically relevant and even selected for in evolution.

A study from the laboratory of Sunney Xie [34•] provides additional support for the 'Poisson' picture as described above in *E. coli*. Normally, it is not possible to observe single molecules of yellow fluorescent protein (YFP) in living cells, because their signal is diluted by rapid diffusion throughout the cytoplasm. This group overcame the challenge and was able to detect single molecules of a fast-maturing variant of YFP by tethering them *in vivo* to the cell membrane, thereby slowing their diffusion [34•]. Coupled with a mathematical analysis, their data then showed that leaky expression from the repressed *lac* promoter was probably the result of single uncorrelated mRNAs that were then translated into geometrically distributed bursts of protein [34•].

Many essential proteins are often present in extremely low numbers. Extrapolating the scaling rule found by Bar-Even *et al.* [13••] for medium-expressing genes, where protein variance is 1200 times the mean level, Baetz and Kærn [35] pointed out that a mean protein level of 200 would correspond to a 500-protein standard deviation, leading presumably to lethal conditions in many cells. In principle, compensatory measures such as negative feedbacks [7] can reduce noise levels below this Poisson limit, but achieving this in practice is quite difficult. Newman *et al.* [14••] find very few proteins whose noise lies below this level, which suggests that such motifs are not often used, at least not in medium-expressing genes. Unfortunately, cellular background autofluorescence prevented both large-scale proteome studies from directly observing these weakly transcribed genes [13••,14••]. In the future this limitation might be overcome using, for example, a gene amplifier circuit [8] or using single-protein techniques [20,34•].

These studies clarify that average expression level is the primary predictor of noise level for much of the proteome. However, although all three implicate mRNA noise as the most likely explanation for the observed scaling, none measures mRNAs directly, leaving some room for doubt as to whether the observed scaling is due entirely to Poisson mRNA fluctuations. As Bar-Even *et al.* [13••] note, infrequent gene activation events might also explain the observed relationship between noise and mean level under certain circumstances (see Figure 1i). This possibility is discussed next.

Direct observations of transcriptional 'bursting'

The 'Poisson' scenario is not the only interpretation consistent with the data. An alternative is that promoters toggle between active and inactive states, reminiscent of what is sometimes called a telegraph process. In this case, lengthy periods of zero transcription are punctuated by the production of numerous mRNAs in quick succession — a transcriptional 'burst' [9,15,27•,36,37]. As in the 'Poisson' scenario, the steady-state mRNA and protein noise might also scale inversely with the mean, but with a proportionality constant determined by the promoter activation kinetics as well as the rates of protein and

mRNA production and the mRNA decay time (see Figure 1). At the protein level, this simple proportionality appears only under certain circumstances: when genes are mostly inactive; when the frequency of activation events is comparable to the rate of protein degradation; and when the average transcriptional burst size is similar between genes [13••]. To confirm if such bursting takes place, the nail in the coffin would be direct real-time observations of individual proteins and mRNA transcripts in living cells [20,26,27•,34•,37]. Several new techniques have recently become available that allow just such observations.

One approach developed by Robert Singer and collaborators allows real-time observation of mRNAs as they are born and diffuse about the cell [38,39]. The technique works by genetically adding a series of hairpin binding sites to the untranslated region of any RNA. These sites allow the docking of GFP molecules fused to the RNA binding protein MS2, which is constitutively expressed. When an mRNA is produced, the GFP molecules rapidly aggregate, forming a visible spot within the cell [38,39]. Golding *et al.* [26••] optimized this technique to allow precise quantification of transcript level produced by a synthetic variant of the *lac* promoter in living *E. coli* cells [40]. The trajectories made visible for the first time what had previously been at most inferred: transcriptional bursts are real. Moreover, they happen even in bacteria, which lack many of the regulatory complexities of eukaryotes [26••]. The study proceeds to quantify this observation, showing that the proportionality constant between the mRNA variance and mean is four, and not unity as would be expected in the absence of bursting (compare Figure 1d and Figure 1e) [26••]. The same technique applied in the amoeba *Dictyostelium discoideum* also revealed transcriptional bursts [37], the first such direct examination in eukaryotes. Unlike the two bacterial studies [26••,34•], *Dictyostelium* cells that produced one burst of RNA were found to be more likely to produce a second burst at a later time, an observation the authors refer to as 'transcriptional memory' [37]. However, this disparity might not signify any general differences between prokaryotes and eukaryotes, but rather simply reflect differences in complexity of the specific promoters examined.

Perhaps the most dramatic example of transcriptional bursting to date comes from a study in mammalian Chinese hamster ovary (CHO) cells. Employing a detection method similar to fluorescent *in situ* hybridization (FISH) where fluorescent DNA probes called molecular beacons are annealed to mRNAs in fixed cells, Raj *et al.* [27•] observed massive transcriptional bursts in both synthetic reporter genes as well as the natural RNA polymerise II gene. Cells actively producing transcripts at the time of fixation showed bright clusters of mRNAs within the nucleus, suggesting that diffusion had not yet taken its course for these transcripts. Other cells showed a smaller number of mRNAs uniformly distributed about the cell [27•]. As a result, the variance of mRNAs between cells was found to be forty times greater that the mean, strongly supporting the 'Telegraph' picture (Figure 1e,f).

In addition, when two distinguishable but otherwise identically regulated reporter genes were integrated in two distinct chromosomes, their bursts were found to be entirely uncorrelated. However, when the reporter genes were moved adjacent to one another at a single chromosomal locus, their mRNA expression patterns became almost fully correlated [27•], consistent with experiments in yeast cells on weakly expressed genes [8]. This finding implicates events upstream of transcription, and not the production of mRNA transcripts per se, as being responsible for the observed noise. The authors simultaneously measured the transcripts of both the reporter gene and the RNA polymerase II gene and found their bursts to be uncorrelated, thereby also ruling out global fluctuations in polymerase concentration as being relevant to the noise of individual genes [27•]. It has been speculated that the entire locus associates with a transcriptionally active region of the nucleus [8]. In possible disagreement with these studies, the genome-wide study of Newman *et al.* [14••] found that genes residing adjacent to one another on a chromosome are as unrelated in noise properties as any two randomly chosen genes. Whether or not these studies are really at odds remains to be seen, because the effects of chromosomal position on noise might simply be more noticeable in weakly expressed genes or in mammalian cells [8,27•].

Conclusions

The relative contributions of mRNA and promoter noise on the overall noise architecture of bacterial and eukaryotic cell populations remains imperfectly known, with genome-wide assays giving a tentative nod to the former whereas single-molecule assays find strong evidence in some cases for the latter. The steady march of technological innovation is likely to settle the matter soon, however. Yet to be combined are high-throughput and single-molecule methods, which would allow single-protein or mRNA resolution for the numerous weakly expressed genes that are most vulnerable to stochastic variation. Likewise, it is only a matter of time before single-mRNA and single-protein assays are combined to allow direct measurement of protein bursts at a known mRNA number. Finally, these combined experiments will need to be done across a variety of organisms to determine whether noise properties can be generalized across phyla.

Acknowledgements

Many thanks to J Tsang, I Golding, J Paulsson, A Raj, J Pedraza and MJT O'Kelly for careful reading and thoughtful feedback on drafts of this manuscript. This work was supported by National Science Foundation (PHY-0548484) and National Institutes of Health (R01-GM077183) grants

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Benzer S: **Induced synthesis of enzymes in bacteria analyzed at the cellular level**. *Biochim Biophys Acta* 1953, **11**:383-395.

2. Maloney PC, Rotman B: **Distribution of suboptimally induces -D-galactosidase in *Escherichia coli*. The enzyme content of individual cells**. *J Mol Biol* 1973, **73**:77-91.

3. Berg OG: **A model for the statistical fluctuations of protein numbers in a microbial population**. *J Theor Biol* 1978, **71**:587-603.

4. Novick A, Weiner M: **Enzyme induction as an all-or-none phenomenon**. *Proc Natl Acad Sci USA* 1957, **43**:553-566.

5. Spudich JL, Koshland DE Jr: **Non-genetic individuality: chance in the single cell**. *Nature* 1976, **262**:467-471.

6. Ko MS, Nakauchi H, Takahashi N: **The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates**. *EMBO J* 1990, **9**:2835-2842.

7. Becskei A, Serrano L: **Engineering stability in gene networks by autoregulation**. *Nature* 2000, **405**:590-593.

8. Becskei A, Kaufmann BB, van Oudenaarden A: **Contributions of low molecule number and chromosomal positioning to stochastic gene expression**. *Nat Genet* 2005, **37**:937-944.

9. Blake WJ, M KA, Cantor CR, Collins JJ: **Noise in eukaryotic gene expression.** *Nature* 2003. **422**:633–637.

10. Colman-Lerner A, Gordon A, Serra E, Chin T, Resnekov O, Endy D, Pesce CG, Brent R: **Regulated cell-to-cell variation in a cell-fate decision system**. *Nature* 2005, **437**:699-706.

11. Pedraza JM, van Oudenaarden A: **Noise propagation in gene networks**. *Science* 2005, **307**:1965-1969.

12. Elowitz MB, Levine AJ, Siggia ED, Swain PS: **Stochastic gene expression in a single cell**. *Science* 2002, **297**:1183-1186.

13. •• Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N: **Noise in protein expression scales with natural protein abundance**. *Nat Genet* 2006, **38**:636-643.

The authors measured expression from 43 genes in each of 11 conditions, resulting in a global trend wherein noise scales inversely with mean. This paper mathematically discusses many possible origins of the observed scaling, including global factors, pathway-specific factors, promoter activation kinetics, and randomness of mRNA production.

14. •• Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS: **Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise**. *Nature* 2006, **441**:840-846.

The largest survey to date of noise across the proteome. The authors argued that stochastic creation and destruction of mRNA describes the bulk of the observed variation. They showed that classes of genes regulated by the same proteins have similar noise levels and that essential genes are often less noisy than inessential ones.

15. Raser JM, O'Shea EK: **Control of stochasticity in eukaryotic gene expression**. *Science* 2004, **304**:1811-1814.

16. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB: **Gene regulation at the single-cell level**. *Science* 2005, **307**:1962-1965.

17. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A: **Regulation of noise in the expression of a single gene**. *Nat Genet* 2002, **31**:69-73.

18. Cookson S, Ostroff N, Pang WL, Volfson D, Hasty J: **Monitoring dynamics of single-cell gene expression over multiple cell cycles**. *Mol Syst Biol* 2005, **1**: 2005 0024.

19. Mettetal JT, Muzzey D, Pedraza JM, Ozbudak EM, van Oudenaarden A: **Predicting stochastic gene expression dynamics in single cells**. *Proc Natl Acad Sci USA* 2006, **103**:7304-7309.

20. Cai L, Friedman N, Xie XS: **Stochastic protein expression in individual cells at the single molecule level**. *Nature* 2006, **440**:358-362.

21. Golding I, Cox EC: **RNA dynamics in live *Escherichia coli* cells**. *Proc Natl Acad Sci USA* 2004, **101**:11310-11315.

22. Rosenfeld N, Perkins TJ, Alon U, Elowitz MB, Swain PS: **A fluctuation method to quantify *in vivo* fluorescence data**. *Biophys J* 2006, **91**:759-766.

23. Suel GM, Garcia-Ojalvo J, Liberman LM, Elowitz MB: **An excitable gene regulatory circuit induces transient cellular differentiation**. *Nature* 2006, **440**:545-550.

24. Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, Alaluf I, Swerdlin N, Perzov N, Danon T, Liron Y *et al.*: **Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins**. *Nat Methods* 2006, **3**:525-531.

25. Volfson D, Marciniak J, Blake WJ, Ostroff N, Tsimring LS, Hasty J: **Origins of extrinsic variability in eukaryotic gene expression**. *Nature* 2006, **439**:861-864.

26. •• Golding I, Paulsson J, Zawilski SM, Cox EC: **Real-time kinetics of gene activity in individual bacteria**. *Cell* 2005, **123**:1025-1036.

This study was the first to directly demonstrate transcriptional bursting by tracking the birth dynamics of individual mRNAs. While in the active state, genes were found to produce mRNAs at a constant rate. The authors also showed that mRNAs partition approximately binomially at cell division, leading to a cell cycle dependence in the correlation between mRNAs and proteins.

27. • Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S: **Stochastic mRNA Synthesis in mammalian cells**. *PLoS Biol* 2006, **4**:e309.

The authors used a FISH-based technique to demonstrate massive transcriptional bursting in mammalian cells. Gene activation events were found to simultaneously affect several genes at a single genomic locus, whereas chromosomally distinct loci switched between active and inactive states independently.

28. Thattai M, van Oudenaarden A: **Intrinsic noise in gene regulatory networks**. *Proc Natl Acad Sci USA* 2001, **98**:8614-8619.

29. Paulsson J, Ehrenberg M: **Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks**. *Phys Rev Lett* 2000, **84**:5447-5450.

30. Paulsson J, Berg OG, Ehrenberg M: **Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation**. *Proc Natl Acad Sci USA* 2000, **97**:7148-7153.

31. Paulsson J, Ehrenberg M: **Noise in a minimal regulatory network: plasmid copy number control**. *Q Rev Biophys* 2001, **34**:1-59.

32. Rigney DR, Schieve WC: **Stochastic model of linear, continuous protein synthesis in bacterial populations**. *J Theor Biol* 1977, **69**:761-766.

33. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB: **Noise minimization in eukaryotic gene expression**. *PLoS Biol* 2004, **2**:e137.

34. • Yu J, Xiao J, Ren X, Lao K, Xie XS: **Probing gene expression in live cells, one protein molecule at a time**. *Science* 2006, **311**:1600-1603.

The authors developed a novel assay allowing the detection of single YFP molecules in living cells. Leaky expression from the repressed *lac* promoter in *Escherichia coli* cells was then shown to result in geometrically distributed protein bursts that were probably the result of single, stochastically produced mRNAs.

35. Baetz K, Kaern M: **Predictable trends in protein noise**. *Nat Genet* 2006, **38**:610-611.

36. Golding I, Cox EC: **Eukaryotic transcription: what does it mean for a gene to be 'on'?** *Curr Biol* 2006, **16**:R371-R373.

37. Chubb JR, Trcek T, Shenoy SM, Singer RH: **Transcriptional pulsing of a developmental gene**. *Curr Biol* 2006, **16**:1018-1025.

38. Levsky JM, Shenoy SM, Pezo RC, Singer RH: **Single-cell gene expression profiling**. *Science* 2002, **297**:836-840.

39. Bertrand E, Chartrand P, Schaefer M, Shenoy SM, Singer RH, Long RM: **Localization of ASH1 mRNA particles in living yeast**. *Mol Cell* 1998, **2**:437-445.

40. Lutz R, Bujard H: **Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements**. *Nucleic Acids Res* 1997, **25**:1203-1210.