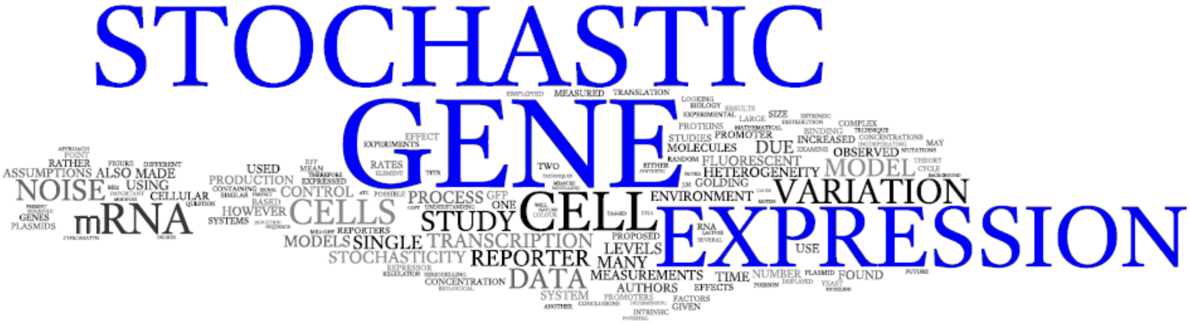


Title: Stochastic Gene Expression

Date: August 19th 2010

Author: E. O Duibhir

Supervisor: Prof. F.C.P. Holstege

Table of Contents

Introduction

Gene expression is the production of macromolecules from a DNA sequence. The macromolecule may be either a protein or a non-coding RNA molecule that is not translated. The gene expression process always involves transcription and may include mRNA processing (eukaryotes), translation (proteins), and further post-translational protein modifications, with each sequential step having the potential to be modified in numerous ways. The regulation of each of these steps essentially controls information flow from DNA to RNA to protein: the central dogma in molecular biology.

A stochastic process involves a factor which cannot be predicted and therefore is best described as a probability distribution rather than having a measured, defined value. The current concept of stochasticity is based on quantum mechanics, where stochastic models have been usefully applied to the understanding and manipulation of subatomic particles. In quantum physics, models are used to represent the uncertainty in either position or momentum of a subatomic particle at any given moment in time. The implications of these theories relate to the nature of the universe and our understanding of reality. Stochastic gene expression is therefore the expression of a macromolecule from a DNA sequence where the process involves a random factor that cannot be determined.

The mathematical modeling techniques developed for subatomic purposes have found many uses in a variety of fields where a large degree of uncertainty is involved (Huffaker, 1998). For example, many independent and unidentified factors that contribute to stock market fluctuations can be conveniently grouped into a stochastic term in a modeling equation to account for the unknown. Stochastic modeling in biology has also been applied at various scales, including populations of molecules, cells and organisms (Horan, 1994).

In this thesis I will argue that while a stochastic element to gene expression may be present, it has not yet been unequivocally demonstrated.

Firstly, the study of stochasticity in relation to gene expression often requires the use of technically challenging procedures involving extremely sensitive measurements liable to numerous perturbations. The microenvironments encountered by an experimental microorganism and the limitations regarding the use of reporter systems, have been either ignored, or underestimated by several investigators. Secondly, although our understanding of biological processes has greatly expanded in the last century, it is still incomplete. This lack of knowledge results in assumptions being made when creating models of complex cellular processes. Deterministic models require detailed knowledge of all the factors involved: concentrations, reaction rates and sub-cellular localizations. Stochastic models circumvent this need by adding a random element to the model to account for the unknown. If the model can then fit the data it has in some cases been assumed that an inherently stochastic process is at work. While the development of sophisticated stochastic models is an extremely useful thought exercise, using such a model in an attempt to prove that a mechanism is inherently stochastic is flawed circular logic.

As modern biology moves from a genomics to a systems biology paradigm the fidelity of information flow from DNA to protein to networks is an important factor to consider. If we are to understand how the cell functions, basing this understanding on solid scientific data is of the utmost importance.

Modelling

Before scrutinising the available data in support of stochastic gene expression, it is useful to begin with an overview of the modelling approaches employed in biological studies of stochasticity. Models form an integral part of these studies and are used in both transformations and interpretations of the data produced. Creating a model to explain a process is a useful thought exercise: the concept of stochasticity in gene expression resulted from a physicist (Erwin Schrödinger) applying a theoretical argument to the concept of gene expression (Rao et al., 2002). The central idea is that a single molecule of DNA is used as a template for several copies of mRNA, which are subsequently used for the production of even more protein molecules. Stochastic events at the level of the single DNA molecule might then be propagated to the level of protein numbers. Models incorporating various degrees of complexity have since been proposed in an attempt to understand and explain variation seen at the mRNA or protein level. Most stochastic modelling approaches are based on the Gillespie algorithm (Gillespie, 1977) or an elaboration of it. The model was originally conceived to allow modelling of coupled chemical reactions by using a pseudorandom number generator (the Monte Carlo method). It was argued by Gillespie that continuous, deterministic models are not a realistic representation of the physical environment, given that numbers of molecules can only exist as integer values and the random nature of Brownian motion affects the microscopic environment. He therefore suggested that a reaction probability per unit time, rather than chemical reaction rates, should be considered when modelling chemical reactions. One of the underlying assumptions in the Gillespie algorithm is that molecules are distributed randomly and uniformly in a given volume.

Approaches that investigate one aspect of gene expression (transcription or translation) and then attempt to model the process, highlight one of the challenges facing gene expression modelling. In one regard investigators have a desire to keep models as simple and robust as possible, while incorporating what are regarded as the most important factors (Bar-Even et al., 2006). More complex models need to make more assumptions than simple ones, in theory leaving them prone to error (Shahrezaei and Swain, 2008). If modelling is used as a theoretical thought experiment to explore the potential of simple interacting factors to produce complex effects, the exercise is useful. However, applying an oversimplified model to complex datasets may result in the dangerous situation of misinterpretation of that data (Pedraza and Paulsson, 2008).

Some of the most useful theoretical modelling experiments are those that question the implications of experimental design, rather than trying to explain results. In one such study the use of protein reporter type was shown, in theory, to have a misleading effect on the results obtained (Zhang et al., 2006). Binary gene induction (see figure 1) is at the centre of many studies exploring stochastic gene expression, given that a switching type mechanism for gene expression, rather than a graded response, could be more susceptible to random fluctuations. The authors ran simulations looking at three reporters; β-galactosidase (β-gal); luciferase and green fluorescent protein (GFP); based on the experimentally measured mRNA and protein half-lives and reporter sensitivity. It was found that as a result of the relatively long protein half-life and low sensitivity of GFP compared to the other reporters, it was more likely to display a graded expression profile. If this model is correct it raises concerns about the use of GFP as a reporter of noise in biological systems. The model also provides an alternative explanation for cellular heterogeneity in response to an externally provided inducer, where a binary response is the result of competition between activators and repressors, rather than stochastic gene expression.

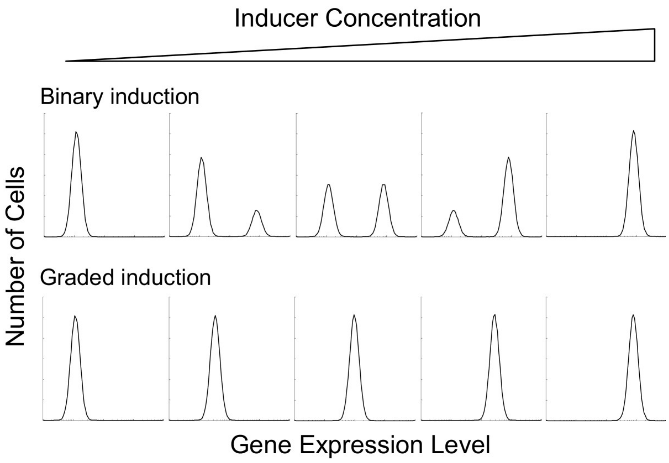


Figure 1: Schematic representation of changes in gene expression due to an inducer. Binary induction involves an on/off state while graded induction displays a dose dependent response. Taken from Zhang et al., 2006.

Another group (Thattai and van Oudenaarden, 2004) modelled stochastic gene expression in a changing environment. The potential for environmental heterogeneity can be overlooked in experimental setups (van Hoek and Hogeweg, 2007). The modelling conditions imposed on the virtual cells were stochastic gene expression and a cycling (but otherwise homogeneous) environment. Under certain conditions it was shown that a dynamically heterogeneous population (as opposed to an initial fixed mixture of static cell types) can achieve higher growth than a homogeneous population. However, the assumptions implicit in the model were non-limiting nutrients and no sources of competition or cooperation existing between cells. This is clearly not a biologically relevant case. It was stated in the text that the changing of phenotype of the cells was intended to be a result of stochastic gene expression. However, this was not implicit in the model and a non-isogenic cell population could in theory give a similar result.

The promoter of the lactose operon is frequently used in expression systems examining gene expression, where synthetic inducers are employed to relieve repression of a reporter gene. Hoek and Hogeweg approached stochastic gene expression from an evolutionary perspective and modelled noise responses to both a synthetic inducer and the natural substrate, using both stochastic and deterministic models (van Hoek and Hogeweg, 2007). They found that bistability (the result of a subpopulation of cells responding to induction in a binary manner) was not present in the lactose induction model and hysteresis, or memory by cells of the previous environment, was only present when synthetic inducers were used. They also highlighted the (potentially invalid) assumptions of environmental homogeneity and isogenic cell populations that are present in most models. By incorporating spatial (environment), genetic (spontaneous mutation), stochastic (gene expression) and cell cycle sources of noise into their models of population heterogeneity, they found that much of the differences between stochastic and deterministic modelling disappeared. At intermediate inducer concentrations gene expression was mostly influenced by spatial heterogeneity. The equations used in these models were necessarily more complex than those used by other groups, but the conditions they examine may be more realistic to the bacterial genome and the microenvironments it encounters, not just in nature but also in a laboratory setting.

Attempting to clarify some of the discrepancies in different studies, Paulsson reanalysed the data from a number of experiments that had used different outputs, normalisation techniques and models to explain their data (Paulsson, 2004). He found that by applying a single model to each of the data sets, a better consensus could be achieved when trying to interpret the results. While resolving some of the issues surrounding stochasticity this study indicates the potential for error, inherent in assumptions made during the modelling process. He also determined that cell-size variation and measurement errors could be attributed to the experimental setup used, which in turn could overshadow organism differences.

A clear advantage of mathematical modelling is the ability to simulate perfect experimental conditions, a case that does not exist in reality. The question of whether stochastic gene expression is the cause, or is a consequence of cellular heterogeneity, may therefore be a 'chicken and egg' argument that cannot be resolved.

Data

Gene expression in single cells

The first experimental evidence that suggested an effect of stochastic gene expression was observed in 1957 (Novick and Weiner, 1957). Novick and Weiner induced β-gal expression in *E. Coli* using Isopropyl-β-D-1-thiogalactopyranoside (IPTG) as an inducer. Rather than assaying the entire culture for β-gal activity, they observed single cells and noticed an 'all or nothing' response to the inducer. This led to speculation that variation in permease levels, a membrane protein increasing IPTG transport into the bacteria and therefore upregulating its own expression, was the cause of this binary effect on the cell population. Stochastic permease expression was suggested as a mechanism, causing cells to respond differently to the concentration of IPTG. Cells expressing enough of the permease protein would rapidly be induced through a feedback loop. The idea was revisited in 1990 (Ko et al., 1990), this time in mammalian cells and using the response of the glucocorticoid receptor to hormones as the basis for the assay.

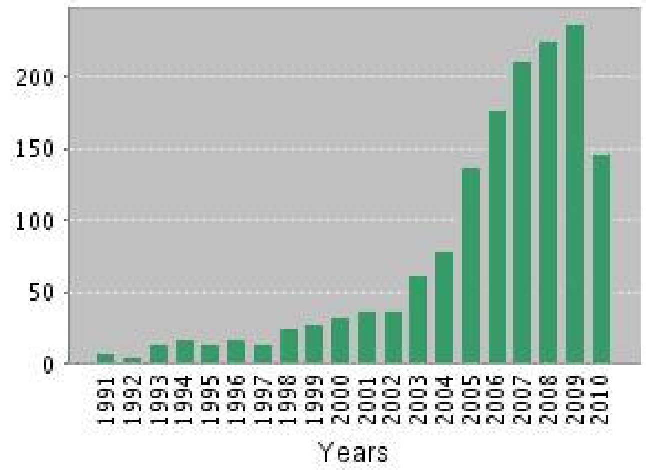


Figure 2: Publications involving stochastic gene expression by year. From ISI Web of Science with conditions: Topic=(stochastic) AND Topic=(gene) AND Topic=(expression)Timespan=All Years, Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH.

Again heterogeneous responses to the inducer were observed and stochastic gene expression was proposed as the culprit. The limited resolution provided by these techniques (both used β-gal expression in their assays) did not allow direct visualization of stochastic events, although ironically they may have made more robust observations due to the reporter system employed, as discussed in the modelling section above (Zhang et al., 2006). In the last ten years there has been a large increase in the number of papers either studying or implicating a stochastic gene expression mechanism in numerous processes (see figure 2). This may be the result of the development of fluorescent reporter proteins, allowing cellular processes to be visualized at a greater resolution. The following sections discuss some of the most influential papers published on the topic.

To examine the effect of auto-regulation on gene expression (Becskei and Serrano, 2000), *E. coli* were transformed with plasmids where the gene product, a green fluorescent protein (GFP), was expressed as a fusion protein linked to the tetracycline repressor (TetR), which then represses its own expression. The amount of fluorescence measured by microscopy in each cell was then taken to represent both the amount of protein produced from the gene and the amount of repressor present in the cell. As controls two other plasmids were constructed. One plasmid expressing the same fusion protein but with a point mutation in the DNA binding region of the TetR protein, that prevented auto-repression and another plasmid that lacked the binding regions for TetR on the gene promoter. These binding regions were replaced by lactose repressor sequences. By varying the levels of anhydrotetracycline (aTc) or IPTG, the maximum concentration of GFP produced by the TetR or lactose systems could be varied respectively. A model was then proposed to explain the effect of auto-regulation on protein variation.

While the aim of the study was not explicitly directed at examining the presence of stochastic gene expression, the authors refer to it during their discussion and many later studies have referenced this paper as an example of the existence of stochastic variability. It does appear from the data that variation is reduced using the auto-regulatory model, however a number of assumptions were made. The expression levels of the unrepressed gene are almost forty times higher than the repressed state. This means that the variances reported are not due to similar expression levels and cannot be compared directly. The effect of production of large amounts of a metabolically useless protein on cell growth is not accounted for. To control for differences in the mean protein concentration the expression level was tuned using various aTc concentrations. Addition of aTc prevents binding of TetR to its response element on the promoter and therefore changes the relationship between GFP as a reporter of gene expression and as a reporter of repression activity, thereby complicating assumptions made in the model construction. Plasmids were used as the vector for gene expression. Variation in plasmid copy number can play a large role in cell to cell variation (Paulsson and Ehrenberg, 2001). High and low copy number plasmids were tested by changing the plasmid origin of replication and assaying for luciferase activity in whole cell lysates, an approach that lacks the resolution to examine cellular heterogeneity due to differential partitioning of plasmids (and a process that has also been proposed to be stochastic in nature (Paulsson and Ehrenberg, 2001)). The data does indicate tighter control of gene expression using auto-repression, however, due to the number of variables unaccounted for in this study, it is impossible to argue for or against a mechanism of stochastic gene expression.

Some of these issues were subsequently addressed by another group in an approach using chromosomally integrated reporters (Ozbudak et al., 2002). Rather than looking at the effects of auto regulation, the expression of GFP was measured using *B. subtilis* strains containing point mutations in either the ribosomal binding site, initiation codon or promoter sequences. The measurement device in this case was a fluorescence-activated cell sorter (FACS). All measurements are reported as noise strength rather than absolute values which, contrary to what the authors declare, make the results more difficult to interpret. It is important to be able to compare the mean expression levels under different experimental conditions. For example, an attempt to control for cellular heterogeneity is made by the use of cell size gating (measured as forward and side scatter) during cell sorting. However, due to the different rates of total protein production measured for each mutant, it is unclear how differences in the ratio of GFP concentration to cell size can be normalised. Unless the increase in cell volume due to growth is perfectly matched by the rate of protein production in each cell, the observed variation in GFP concentration could be attributed to potential dilution of GFP caused by variations in cell size. In other words the production rates could be constant and the fluctuations may arise from variations in cell size.

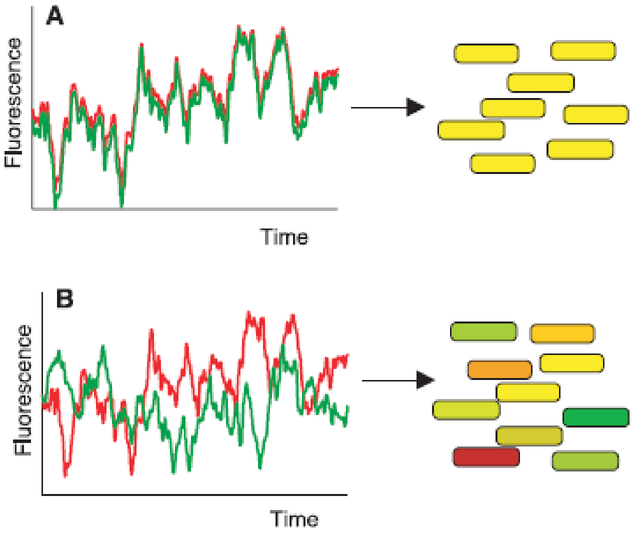


Figure 3: Cartoon of extrinsic (A) versus intrinsic (B) noise measured with two differently coloured fluorescent protein reporters. Taken from Elowitz et al., 2002.

Shortly after the previous study a more robust setup was employed (Elowitz et al., 2002), with the explicit intention of examining stochastic gene expression in a single cell. This paper laid the groundwork and defined the terminology to be used in many future investigations of stochasticity. In the experimental setup two distinguishable fluorescent proteins were stably integrated into bacteria at loci equidistant from the chromosomal origin of replication, both under the control of identical promoters. This provided an internal calibration measure for the variation in expression between individual cells. The authors refer to the difference in fluorescence in a single cell as intrinsic noise, while the difference between cells is termed extrinsic noise. Noise refers to the coefficient of variation, which is the standard deviation divided by the mean. A cartoon representation of these different noise types is shown in figure 3. Under the control of strong constitutively active promoters the cells exhibited little difference in intrinsic noise. The expression of the reporters was then examined under the control of the lactose repressor, with varying concentrations of IPTG added to the medium to induce expression. It was found that at low expression levels

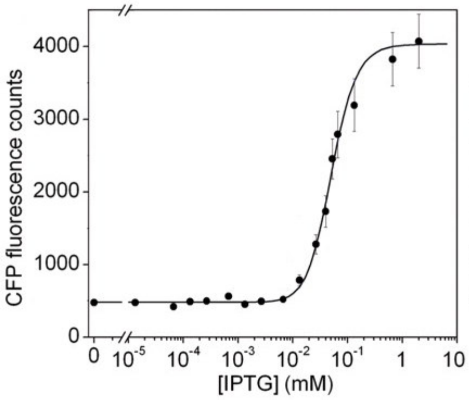


Figure 4: Induction of CFP expression (under control of the lactose repressor) by IPTG. Line is fitting of a Hill function to the data points. Taken from Pedraza and van Oudenaarden, 2005.

extrinsic noise was relatively high. The noise passed through a maximum at an intermediate IPTG concentration and was lowest under full expression conditions. This is as would be expected given the amplification of the effects of cellular heterogeneity at intermediate concentrations of a repressor (see figure 4). At an intermediate concentration the change from repression to expression of a protein hinges about an inflection point that depends on the IPTG concentration in the medium. This process can be modelled using a Hill function, which describes cooperative binding of macromolecules. The steepness of the curve indicates the sensitivity of response to the stimulus. Maintaining a concentration about this inflection point should give the greatest cellular heterogeneity with respect to reporter expression. This had previously been shown experimentally, where a rheostat type gene response could be converted to an on/off control by use of competing activators and repressors (Rossi et al., 2000). The location of the gene producing the lac-repressor protein greatly affected extrinsic noise, with a plasmid, rather than chromosomal expression system, causing greater variation between cells. Bacteria with a background mutation in recA, a protein involved in DNA replication, also displayed increased gene expression variation. Such a mutation induces errors in replication fork elongation, explaining the increased variation in terms of the chromosomal copy number of reporter.

The two colour method developed by Elowitz and colleagues was an important advance in the study of stochastic gene expression because factors affecting protein production external to a cell (intracellular heterogeneity), can be separated from factors inside a cell (intercellular heterogeneity). However, conclusions were drawn based on the assumption of a perfectly isogenic population of cells. It has been shown that *E. coli* can optimise expression of the lacose operon in less than a hundred generations (Dekel and Alon, 2005), which raises questions about the rate of mutation in bacteria and its potential for unexplored effects on gene expression. A perfectly mixed homogenous environment is also assumed in the modelling equations. Neither of these conditions can be guaranteed, as they would both require a perfect experimental system, as discussed in the modelling section above. The application of the two colour method also makes a third critical assumption: that the intercellular environment is a well mixed homogenous environment in itself, while this is not the case (Golding and Cox, 2006).This technique does provide a natural calibration for cell size and addresses many of the confounding effects of heterogeneity that plague later studies of variation in gene expression. It is surprising that future studies often choose to disregard the benefits of this approach.

The study of stochastic gene expression soon moved from prokaryotes to eukaryotes with the two colour system being employed in the budding yeast *S. cerevisiae* (Raser and O'Shea, 2004).

Diploid yeast strains were constructed that expressed CFP and YFP from the same locus on homologous chromosomes, under the control of identical promoters (see figure 5). Importantly in this experiment, the reduction of extrinsic noise by correcting for cell size and cell cycle stage was examined. It was found that approximately a quarter of the extrinsic noise was removed using these corrections. This has implications for future genome wide experiments that relied on a single colour fluorescent reporter protein (namely Newman et al., 2006, discussed in a later section). The variation in protein expression was measured for a variety of promoters (GAL1, PHO84 and PHO5), under conditions that resulted in increased reporter expression (employing either galactose induction or phosphate starvation). Intrinsic protein variation was then measured when using different promoters and found to be promoter dependent. Stochastic models incorporating various reaction rates were subsequently evaluated by their ability to account for the proposed source of stochastic gene expression. These included chromatin remodelling, transcription, translation and mRNA/protein degradation rates. It was found that a model incorporating a large degree of variability due to the chromatin remodelling rate best fitted the experimental data. To determine if the conclusion drawn from the modelling exercise had any value, the reporter system was transferred into a genetic background that had point mutations in essential chromatin remodelling components of the SWI/SNF complex. In these mutant yeast intrinsic noise was increased. Mutations in the TATA box of the PHO5 promoter were then evaluated, as TATA box promoter binding is not believed to be dependent on chromatin remodelling. This decreased both maximal protein expression and intrinsic protein variation.

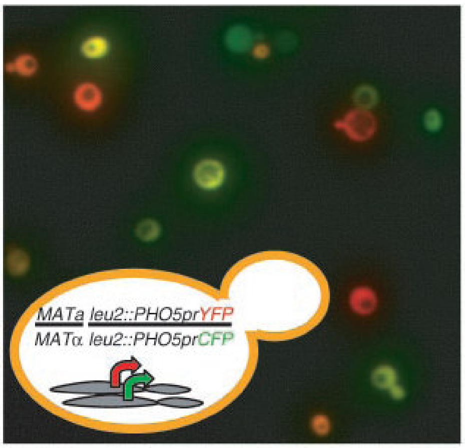
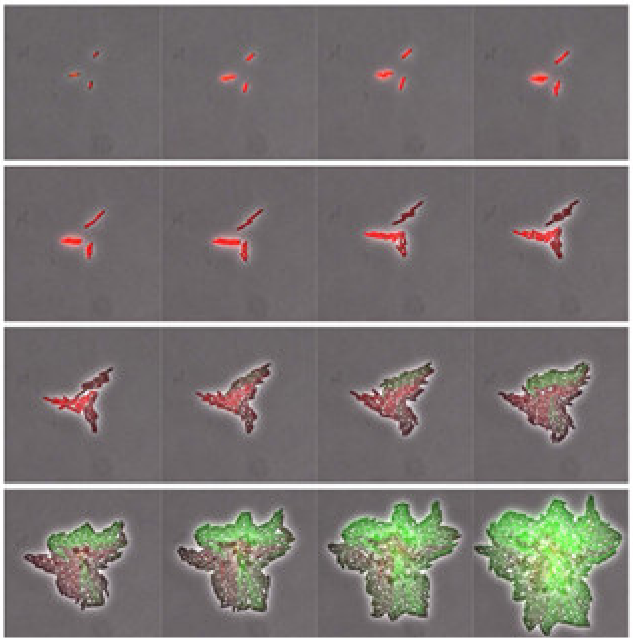


Figure 5: Budding yeast expressing different levels of reporter proteins and inset cartoon of expression system. Taken from Raser and O'Shea, 2004.

This study was well conceived, straightforward, and extensively validated. The authors experimentally tested for the independence of reporters, by showing that no differences were observed in the fluorescence distributions, based on Kolmogorov-Smirnov tests (a statistical test allowing the comparison of distributions, to determine the likelihood that they both came from the same distribution). However, there remain some unresolved conjectures that were unaddressed by the authors. As in the previous study an isogenic population of cells is assumed. Although cells were grown from a single colony, mutations may have occurred in or around areas of the genome where one of the reporters is located (Lynch et al., 2008). While unlikely, this could affect expression levels. The authors did observe that less than 0.1% of cells only expressed one of the proteins, and these cells were excluded from the analysis. While this may seem like a small number, it draws attention to a proportion of cells that were incapable of protein production at one locus (presumably by mutation, deletion or silencing). It is therefore plausible that protein expression levels could also have been affected by similar mechanisms. A well mixed intercellular environment is a precondition for the simple modelling equations proposed and assumes that both genes have access to exactly the same transcription machinery in the nucleus. As mentioned previously, fitting a stochastic model to data does not prove that a stochastic process is at work. The authors seem to concede this point through further verification of their model in a chromatin remodelling compromised background, however, the gross effects on the regulation of many genes in the cell cannot be accounted for. Indeed, deletion of the SWI/SNF complex is lethal (Steger et al., 2003), forcing the use of strains containing point mutations. In the complex intercellular environment where many factors depend on one another, it is difficult to attribute changes in the variations of measurements to one underlying factor, especially when viewed through the forced perspective of a mathematical abstraction involving numerous assumptions. The model proposed disregards the many regulatory steps preceding fluorescent protein measurements that include aspects of transcription, mRNA processing and export from the nucleus, translation, protein folding and chromophore maturation (Golding et al., 2005).

Expression of several genes in a single cell

Based on the conclusions of previous experiments looking at gene expression variation, two studies (Rosenfeld et al., 2005; Pedraza and van Oudenaarden, 2005), published side by side in the same journal, both investigated the propagation of noise in simple bacterial transcriptional networks. While they both made great progress in the understanding of transcription cascades, their contribution to the question of stochastic gene expression is ambiguous.



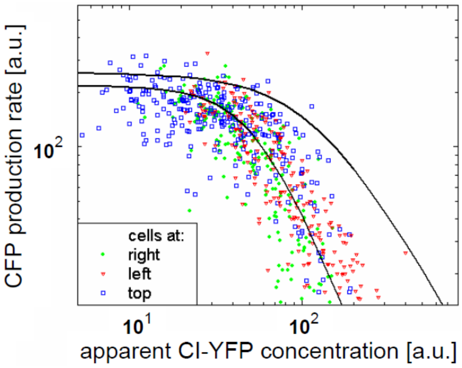

CFP production rate [a.u.]
10^2
cells at:
right
left
top
10^1
10^2
apparent CI-YFP concentration [a.u.]

Figure 6: Comparison of calculated GRFs for three cell colonies based on calculated production rates and concentrations after model fitting. Lack of difference between each cell is given as evidence for lack of effect from cell microenvironment. Taken from supplemental information in Rosenfeld et al., 2005.

The first study (Rosenfeld et al., 2005) was novel in that it used time lapse microscopy. These real-time measurements allowed the investigators to examine the effect of fluorescent repressor dilution (in red) on fluorescent protein reporter expression (green) in single cells. The potential contribution of cellular heterogeneity due to the microenvironment was considered, as was cell cycle phase and size. The authors made many complex transformations to their data employing various models, which makes it difficult to draw convincing conclusions about stochastic gene expression from what is presented. For example: based on a binomial segregation error model, fluorescent data was first converted to the number of molecules per cell; production rates were then calculated, after correcting for cell cycle phase, as production rates were found to approximately double the closer the cell got to division; production rates were fitted by Hill functions (as repression was expected to be cooperative), and a value termed the gene regulation function (GRF) was calculated. This data was taken from three closely positioned cells that formed colonies eventually merging together (see figure 6, left). Based on a comparison of calculated GRF values for each of the cells in each of the colonies, it was stated that the local microenvironment had little detectable effect on GRFs (see figure 6, right), and therefore, there was no effect of microenvironment on the experiment. This conclusion is inconclusive given the conceded cumulative errors in each model fitting and data normalisation stage of the GRF calculation.

The second study (Pedraza and van Oudenaarden, 2005) was novel in terms of the sheer complexity of its experimental setup (partly displayed in figure 7). Three fluorescent reporters were employed, with two being dependent on the expression of repressor proteins upstream in the network. Given the location of the reporters on plasmids with variable copy numbers, this study is vulnerable to the variations in expression encountered previously (Becskei and Serrano, 2000). Although in theory the system employed a two colour method to control for this variation, all measurements are normalised to RFP expression (under control of a different promoter), thereby effectively making this a one colour study with relative mean protein concentrations unable to be otherwise accounted for. Unexpected correlations between the transmission of noise from one gene to another was noted, which is largely attributed to noise as a result of stochastic gene expression.

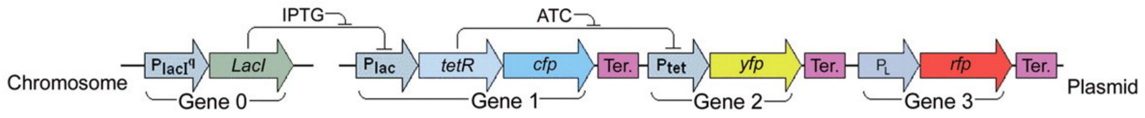


Figure 7: Experimental setup used in Pedraza and van Oudenaarden, 2005.

Each of the studies discussed thus far have made significant steps towards revealing potential stochastic factors in gene expression. These contributions were either conceptual, technical or provided a mathematical framework for further analysis. Unfortunately, the role of stochasticity in gene expression was essentially an assumed fact, with the aim of these experiments being more focused on protein variability or propagation of noise in synthetic transcription networks, rather than looking at the process of gene expression itself.

Gene expression kinetics

The first investigation to directly scrutinise the question of stochastic gene expression incorporated real-time monitoring of mRNA and protein concentrations in a living bacterium (Golding et al., 2005). This was achieved by the use of an ingenious technique, whereby a specific RNA nucleotide sequence motif is recognised by a phage coat protein (Valegard et al., 1997). A gene was expressed containing 96 of these repeating motifs in its mRNA, in cells where the phage coat protein (MS2) was fused to GFP (MS2-GFP) and had been previously expressed. This technique was originally developed to monitor mRNA localisation in budding yeast (Bertrand et al., 1998), but to visualise mRNA production in real-time, Golding and colleagues were limited to cells lacking a nucleus (i.e. plasmid expression in bacteria) as the MS2-GFP protein would not have access to the nucleus of a yeast cell. The experiment required extensive validation and tuning of the system. If the unbound MS2-GFP levels were too high, a background fluorescence signal would obscure the tagging of an mRNA molecule. If the levels were too low, mRNA molecules would not be saturated by reporter and the measurements could not have been trusted. Production of the MS2-GFP protein was under the control of a tetracycline response element, allowing control of expression levels by aTc, while the modified mRNA was transcribed after induction/de-repression of a dual arabinose/lactose controlled promoter. The protein product of the tagged mRNA was itself a fluorescent reporter protein: RFP. This system allowed the direct visualisation of both a single mRNA and the concentration of its protein product at the same time (see figure 8).

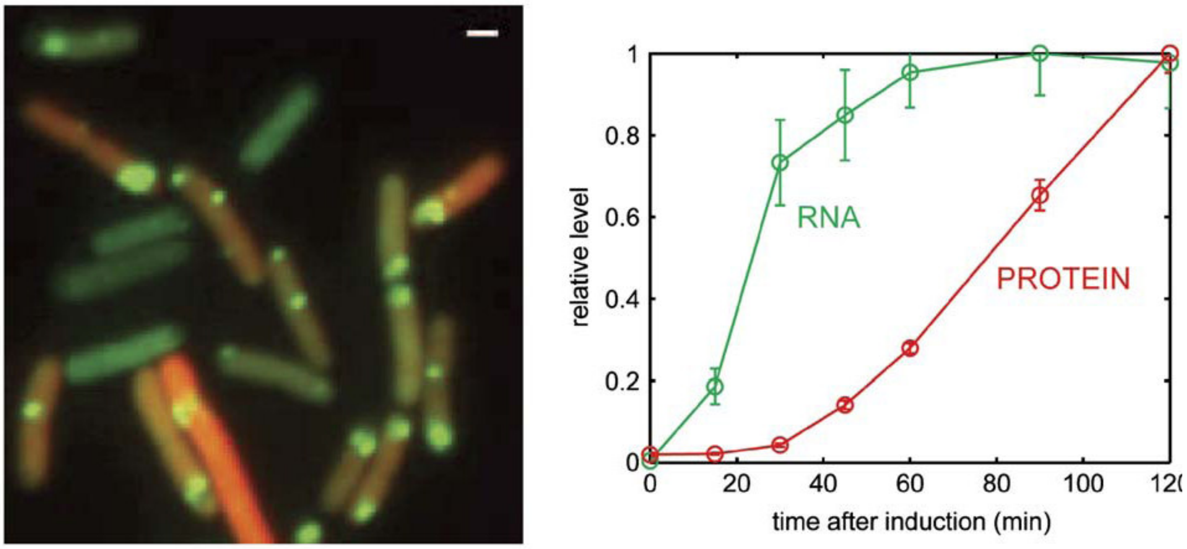


Figure 8: Image of tagged mRNA (MS2-GFP in green) and protein produced from it (RFP in red) in bacteria. Average expression of both based on multiple cell recordings are graphed on the right. Taken from Golding et al., 2005.

After tuning of the system the induction of single mRNA molecules were detected as quanta, allowing larger spots of fluorescence to be subdivided into discrete mRNA molecules. Over several generations the signal from a single mRNA reduced gradually. The authors interpreted this as the mRNA being 'chewed up', resulting in MS2-GFP molecules disassociating. Expression of RFP using native mRNA (lacking the 96 MS2 binding motifs), displayed protein concentrations similar to levels when using the tagged construct, although this was not the case when the motifs were placed upstream of the RFP coding sequence. Single molecule counts of mRNA agreed with whole population qPCR measurements. These measurements indirectly indicated that transcription and translation rates were not largely influenced by inclusion of the MS2 sequence, as it was used in further experiments. A control similar to a dye swap in microarray experiments was also performed, where the GFP and RFP proteins were interchanged on each of the constructs and displayed no differences. Bacterial cells were then followed over several generations, with production of new mRNA molecules and the concentration of RFP in each cell recorded every 30 seconds (displayed in figure 9).

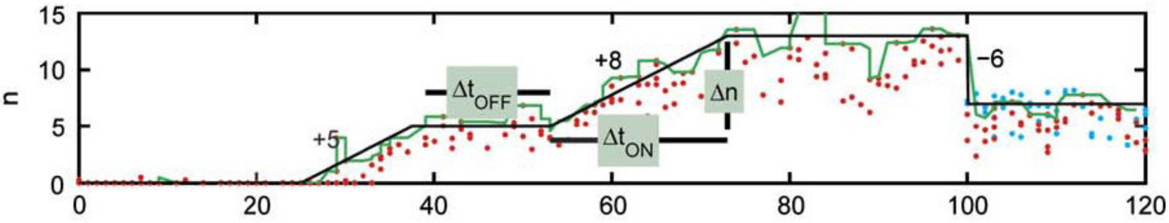


Figure 9: Counting of mRNA molecules (n, y-axis) through time (x-axis, minutes) in a single cell. Δt_{off} and Δt_{on} refer to transcription off and on while Δn refers to the change in number of mRNA molecules. The drop in number at 100 minutes is due to a cell division event. From Golding et al., 2005.

Transcription of mRNA was found to occur in bursts, a theory that had been proposed 27 years earlier (Berg, 1978), based on purely theoretical grounds. The authors then attempted to determine if transcription is a stochastic process by statistically analysing their data. They state that in an idealised situation, if a process is random, it should have a probability distribution that is Poisson, where the variance is equal to the mean. This distribution is based on the likelihood of an event not occurring (or having zero events), which decays exponentially over time (or put another way the chance of an event occurring increases over time). The hypothesis was then tested, by examining if the data displayed a Poisson distribution. The number of cells induced after a given time was measured as the fraction of cells containing a tagged mRNA transcript. The test failed, indicating that induction of transcription was not random. This could be interpreted as a result confirming either that cells are not isogenic, or are not in a perfectly homogenous environment, however the authors do not indicate that they considered these possibilities. A more direct test of the Poisson hypothesis could in theory be achieved if the time between transcription bursts in a single cell was evaluated, alleviating the potentially confounding genetic and environmental effects. This test was performed and the interval times for Δt_{off} were found by the investigators to have a Poisson distribution, while Δt_{on} was also found to be Poissonian, but with a proportionality constant equal to the average mRNA burst size. To explain this finding a modified Poisson process was proposed, incorporating burst-like transcription into their model.

In terms of stochasticity, this study (Golding et al., 2005) emerges as the most detailed analysis of gene expression kinetics to date and is strongly suggestive of a stochastic element, to at least prokaryotic gene expression. However, a number of assumptions are present in the interpretation of the data. As the cell cycle progresses, corrections based on cell size and cell divisions accumulate as errors in measurements, reducing the confidence in mRNA counting as the transcripts themselves accumulate and are divided between daughter cells. The number of data points required to test distribution shapes reliably, means that measurements both before and after cell division must have been incorporated. While reporter protein levels were similar for tagged and untagged mRNA, interactions between mature transcripts and nascent mRNA, in the process of being actively transcribed, could not be ruled out (Golding and Cox, 2004). This raises the question of how representative the observed process is, compared to normal transcription activities in the cell. It was noted by the authors that the added complication of including mRNA degradation rates to their model was alleviated, due to the increased stability of mRNAs containing the repeated MS2 motifs.

Golding and colleagues used reporters that were located on plasmids (Golding et al., 2005). This introduces the possibility of variation due to replication and partitioning of plasmids, once measurements are made beyond a single cell division (Paulsson and Ehrenberg, 2001). The authors presented a simplified stochastic model of gene expression that could account for the observed data. However, a deterministic model of gene expression at different cell cycle time points could, in theory, resemble a stochastic model looking at the entire cell cycle (Berg, 1978). Unfortunately, the difficulty in measuring all the factors required for a deterministic model precludes this test. Models of delay induced oscillations have been proposed (Bratsun et al., 2005) that can empirically resemble stochastic processes and could be used as an alternative to interpret these results (delays from RNA polymerase pausing for example). An extremely interesting prospect would be the implementation of the two colour system to this real-time experiment, analysing the transcription kinetics of two reporter genes simultaneously in the same cell. Comparisons of rates, timing and intercellular location might help to elucidate the determining factors in intrinsic versus extrinsic noise previously observed by Elowitz and colleagues (Elowitz et al., 2002).

Further single cell and single molecule characterisations of translation were made possible shortly after the Golding study by use of a fluorescent reporter protein fused to a membrane targeting peptide (Yu et al., 2006). The technique was used to examine the number of permease proteins required for induction of the lactose operon in *E. Coli* (Choi et al., 2008), one of the classical systems where stochasticity has been invoked (Novick and Weiner, 1957). The simultaneous monitoring of single mRNA molecules (Golding and Cox, 2004) and their protein products (Yu et al., 2006) was then made possible and may have improved the models of gene expression, incorporating potential variation due to translation. Unfortunately, the techniques were not combined and mRNA quantities in the Choi study were estimated by cell population qPCR measurements (Choi et al., 2008), leading to assumptions regarding mRNA number in the model proposed. A combination of these techniques may be a useful approach for future determination of gene expression kinetics in bacteria.

Using a concept somewhat similar to the MS2 reporter system, the variations in mRNA transcript number were recorded in mammalian cells (Raj et al., 2006). This was achieved by stable integration of a fluorescent protein reporter gene into Chinese Hamster Ovary (CHO) cells that contained 32 probe binding sites in the 3’ untranslated region of its mRNA transcript. The cells were stimulated to produce the reporter mRNA by using the tetracycline on (TetO) system (the CHO cells constitutively produced tTa), allowing expression to be tuned by adding doxycycline to the medium. Cells were then fixed, permeablized and probed for mRNA molecules by RNA fluorescent *in situ* hybridization (RNA-FISH). Bright spots of fluorescence could be observed in the nucleus, presumably the sites of transcription factories (Sutherland and Bickmore, 2009), while the cytoplasmic levels of mRNA varied wildly from cell to cell. Reporter genes incorporated at the same locus, but containing different probe sequences, displayed correlated expression, while genes at distant loci did not, confirming a well documented link between chromosomal location and gene activity (Sexton et al., 2007). Reporter proteins with shortened half-lives were expressed by incorporating C-terminal amino acid sequences, actively targeting them for destruction. After these modifications the correlation between protein and mRNA levels was found to approximately double. Fortuitously, due to a repeating motif in the mRNA of the large subunit of RNA polymerase II, its native mRNA could be observed using a modified probe. A large degree of variation was observed for this mRNA, with no correlation found between native RNA polymerase mRNA and reporter mRNA levels. A stochastic model was then proposed to explain the observed data.

Correlations between active transcription sites and the number of mRNA molecules counted per cell are interpreted by the authors (Raj et al., 2006) as evidence of burst-like transcription in these experiments. Given the requirement to fix cells for the RNA-FISH procedure, a snapshot of each cell is observed, with proof of burst-like transcription requiring real-time measurements. More likely sources of expression variation are chromosomal abnormalities (Derouazi et al., 2006) and/or environmental heterogeneity. Adherent mammalian cells grown in culture are known for their diverse morphological phenotypes and varied response to stimuli, with a cell's position relative to its neighbour affecting, for example, its susceptibility to viral infection (Snijder et al., 2009 and accompanying comment in refs.).

The ambitious conclusions and title of this study (Raj et al., 2006) are not justified by the data presented. However, a reasonable point is made by the authors regarding the relationship between standard deviation and mean values and their use as a measure of expression noise: the shape of a histogram is sometimes more important than the ratio of standard deviation to mean.

Expression of many genes in many cells

Studies of stochastic gene expression in single cells soon moved to the examination of expression of many genes in many cells, while retaining single cell resolution. The stated aim of several investigations was the study of biological noise, a term that does not address the source of the noise measured, other than it being biological rather than technical in nature. In this case cellular variation, rather than equipment noise, is being measured, with the reporter signal being above the noise level of the measuring instrument. While it is not possible to infer stochasticity directly from the results, the concept of stochastic gene expression has made a significant contribution to the interpretation of the data generated by these experiments.

In terms of genome wide coverage, the most comprehensive study of protein variation was performed in *S. cerevisiae* by Newman and colleagues (Newman et al., 2006). The study used a collection of haploid yeast strains expressing GFP fusion proteins from their native loci (Huh et al., 2003). The collection was originally constructed to examine the sub-cellular localisation of proteins and covered approximately three quarters of the genome. The experimental setup used by Newman and colleagues was highly automated, with yeast grown from single colonies in agitated 2 ml deep 96 well plates and each well containing a glass bead to improve mixing. The measurements were recorded by FACS and each of the strains were grown in either rich (YEPD) or minimal (SD) media to compare the expression patterns of proteins under different growth conditions. More than 60% of the yeast collection could be measured above autofluorescence (due to fluorescent small molecule metabolites), resulting in fusion protein concentration measurements for approximately half the yeast genome. It was found that GFP did not interfere with the recognition and destruction of proteins through ubiquitination, however this was confirmed for only two yeast strains (less than 0.1% of strains measured). Good repeatability was achieved on different days using different batches of medium. The difference between intrinsic and extrinsic contributions to variation in protein levels was also assessed, by examining diploid yeast incorporating a red fluorescent protein at the same locus. Interestingly, there were large differences in the relative contributions of noise type between the four proteins measured in this fashion, as shown in figure 10. In a test for reporter independence Rps25A failed: the expression of RFP was therefore affecting GFP expression. This was measured by comparing the mean GFP signal from single and dual colour measurements. ATP4 also displayed a 15% difference from average, once the RFP protein was co-expressed, but was deemed to pass the test. In terms of reporter equivalence, datasets generated separately with each fluorophore showed differences based on Kolmogorov-Smirnov tests. This was justified as being due to the sensitivity of FACS and a reduction in the number of data points (to a level commonly used in microscopic experiments) then displayed no statistical difference.

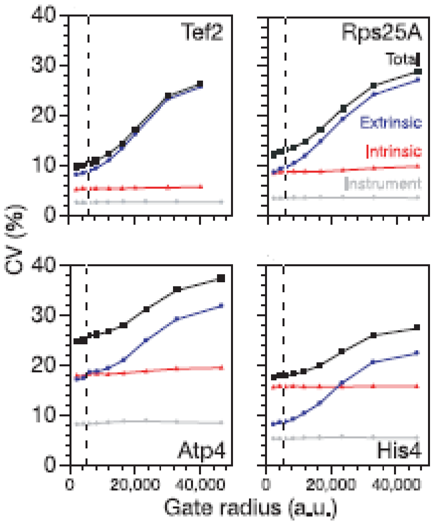


Figure 10: Intrinsic versus extrinsic noise for the four fusion proteins examined by FACS using the two colour method in the study. Dotted vertical line displays setting of the FACS scatter gate for further analysis. Taken from Newman et al., 2006.

From the perspective of stochastic gene expression, the dual colour data is the most relevant data provided by the study, however, the authors use the data as a justification for even smaller gating of the cells used in the analysis, rather than exploring the implications further (although it can clearly be seen in figure 10 that the contribution of intrinsic and extrinsic noise varies depending on the gene observed, also below the gate line chosen). After FACS analysis cell measurements were gated, firstly on forward and side scatter, to reduce the variation due to cell size and granularity and secondly, to reduce the variation in intensities of GFP. This resulted in the data from approximately 1% of all cells examined going forward for further analysis. The logic of this second data trimming step seems counterintuitive in a study aimed at revealing the architecture of biological noise, while caveats regarding corrections based on cell size and the reduction of extrinsic noise had been noted previously (Raser and O'Shea, 2004).

After ensuring that instrument noise was not making significant contributions to the noise measured in GFP fluorescence readings, the authors then turned their attention towards the potential sources of biological noise. The data was subjected to a nonlinear transformation using a running median of coefficient of variation values, resembling a LOWESS normalisation. Each protein was then given a DM value (not defined in the original paper or the supplementary text). The use of these values is reported to be less affected by mean protein abundance or intracellular differences. The DM values are then correlated to Gene Ontology (GO) terms with one of the most prominent correlations found relating to the mode of transcriptional regulation. The authors suggested that transcriptional regulation is a major source of protein variation and that greater protein abundance generally reduces variation. This was proposed to be due to stochastic processes in transcription, supported by references to some of the studies discussed previously (Ozbudak et al., 2002; Elowitz et al., 2002; Raser and O'Shea, 2004).

While making corrections for auto-fluorescence to their data, it was noticed that heterogeneity in auto-fluorescence readings were also present. This could indicate the presence of variation in the metabolic state of the cells being monitored (Billinton and Knight, 2001). Metabolic oscillations in yeast are known to occur at time scales shorter than the cell cycle (Richard, 2003), possibly making corrections based on cell size even more unreliable. Expression of metabolic genes was previously found to be modular (Ihmels et al., 2004), which may itself contribute to cellular heterogeneity. The GFP variant (S65T) used in creation of the original yeast collection was chosen for its fast maturation and protein stability (Huh et al., 2003). This variant of GFP was found to be two fold brighter at a pH range of 6-7 (EGFP, figure 11, (Llopis et al., 1998)). Improper protein folding has also been noted at different temperatures, an effect that may be pronounced in deep 96 well plates.

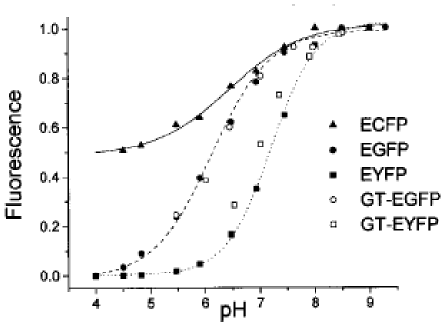


Figure 11: Dependence of fluorescence intensity on protein reporter mutant type and pH. Taken from Llopis et al., 1998.

In theory these effects should be the same for all the different yeast strains examined, assuming these fusion-proteins do not affect metabolism. However, the growth kinetics of all viable yeast strains cannot be determined from this data, as cells were grown to a similar optical density and then sampled, masking potential differences in growth kinetics. Comparison of mean fluorescence levels may also be affected by sub-cellular localisation, indeed GFP variants have previously been used to study pH variations in sub-cellular compartments (Llopis et al., 1998). Different cellular compartments display widely varied pH properties. In mammalian cells the cytoplasm, mitochondria and Golgi measured pH values of 7.3, 7.9 and 6.6 respectively (Llopis et al., 1998). Comparing noise data normalised by protein abundance may therefore be prone to errors and biases related to cellular location. When these values were then correlated to GO terms, already proven to correlate well with cellular localisation (Huh et al., 2003) the p-values calculated for the proposed correlations are meaningless. Over 44% of fusion-proteins in the yeast collection were found to localize to specific compartments, rather than being free in the cytoplasm.

Newman and colleagues also compared their data to data generated from un-gated, un-tagged yeast strains in microarray experiments. They found a good correlation between their FACS data and microarray data, when changes in expression were observed depending on whether yeast were grown in either YEPD or SD medium, but noted that some changes were not picked up in the microarray data. Another study using the same yeast collection and looking at the correlation between changes in the expression data of microarrays compared to FACS based GFP-fusion fluorescence, proposed post transcriptional regulation as a potential source of discrepancy between the two techniques (Lee et al., 2007). While this may be true in some cases (and was shown by comparing Western blotting and qPCR data for several examples (Newman et al., 2006)), another (unlikely) possibility is a change in cellular localisation of a protein depending on the medium type used (YEPD or SD). The corresponding change in local pH could then change the fluorescence levels measured from the fusion protein. This possibility could have been tested microscopically. Newman did find what is described as an 'unexpected correlation' between noise and sub-cellular localisation, with low noise proteins enriched in the golgi and high noise proteins enriched in mitochondria. The cause of this was postulated to be due to unequal partitioning of organelles during cell division and was supported by an experiment using yeast with a gene knocked out that is known to cause unequal partitioning of peroxisomes (Inp1). An increase in noise in peroxisome related fusion proteins was then observed, although this result is questionable, given that it was noted in the original collection analysis (Huh et al., 2003) that proteins targeted to the peroxisome had trouble finding their correct location (due to disruption of the a C-terminus localisation signal by C-terminal GFP fusion). Another explanation may be that the variation in fluorescent signal is dependent on the local pH encountered by GFP molecules, which fluctuates according to the ultradian metabolic phase of an organelle. This could explain the increased noise noted at the mitochondria, a highly metabolically active organelle, and might also help to explain some of the discrepancies between FACS and western blotting data. The exact sources of noise (and there are probably many) found in these experiments is difficult to determine. The frustrating point here is the assumption that stochastic gene expression may have made a significant contribution to biological noise, based on simplistic models, while stochasticity is still unproven in eukaryotic cell studies.

A different approach to the data analysis, aimed at identifying causes of intercellular heterogeneity, may have been more useful: a source of variation frequently ignored in studies directed towards proving stochasticity. The authors did notice a bimodal distribution to protein levels in their data. These distributions were attributed to cell cycle variations. A CDC score (similarity to cdc15 and cdc28 gene expression cycles) was previously proposed based on time course microarray experiments performed after cell cycle synchronisation (Spellman et al., 1998). The CDC score is a measure of the correlation in expression cycle with known cell cycle markers. It was found that for abundant proteins there was a good correlation between bimodal fluorescence intensity distributions and CDC score (Newman et al., 2006; see figure 12). Trimming of the raw data with the gating procedures described previously removed this correlation. It would be interesting to determine which of the proteins not related to the cell cycle displayed a bimodal pattern, as these could be indicative of subpopulations of cells

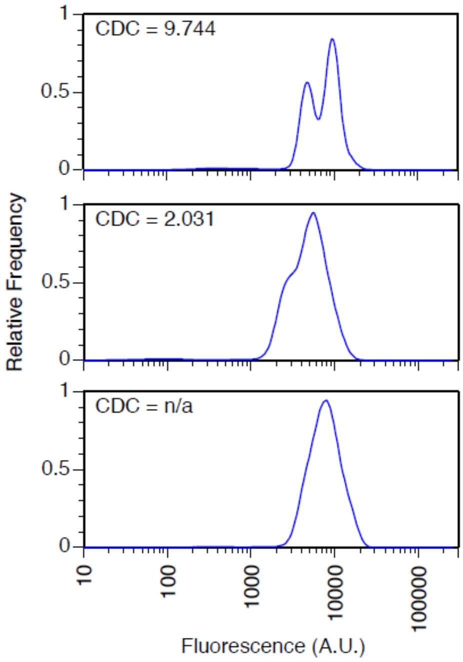


Figure 12: Fluorescence intensity distributions and CDC score. Hhf2-GFP (top panel), Cwp2-GFP (middle panel) and Ssb2-GFP (bottom panel). Taken from supplementary information in Newman et al., 2006.

present in the culture. The authors did note that many of their intensity histograms had long tails that were also removed due to the gating procedure employed. Yeast displaying either high or low mean levels of fluorescence were physically sorted into two different groups and their expression levels were found to be maintained when re-measured. This could be due to cellular replicative aging, with daughter cells receiving newer proteins than their mother (Eldakak et al., 2010). Intensity histograms of these proteins would also be expected to display long tails.

A similar study performed with the same yeast collection did look at the effects of environment on gene expression (Bar-Even et al., 2006). In this setup cells expressing 38 GFP fusion proteins were analysed by FACS, after perturbation of growth by transferring the yeast to one of 11 different environmental conditions. The genes were subdivided into four different co-expressed transcription modules (stress, proteasome, ergosterol and rRNA processing), with the environmental perturbations including eight stress inducing conditions and three relaxation conditions. After perturbation the cells were analysed at six time points (every 30 minutes). The greatest correlation was found between mean protein abundance and noise levels, with increased expression leading to decreased noise. The results were then interrogated using stochastic models of gene expression and the conclusion was drawn that this abundance/noise dependence was the result of stochastic birth and decay of mRNA molecules.

The dependence of noise on mean expression levels had been noted previously (Becskei et al., 2005) in an attempt to correlate chromosomal position to noise levels (using an expression system 'amplifying' noise through an expression cascade). While the perturbation study (Bar-Even et al., 2006) suffers from many of the same potential ambiguities that were noted for the previous study (Newman et al., 2006), and possibly even more so if metabolic fluctuations are an important source of noise, the gating procedures used were not as stringent, giving a better view of cellular responses to different environments (see figure 13).

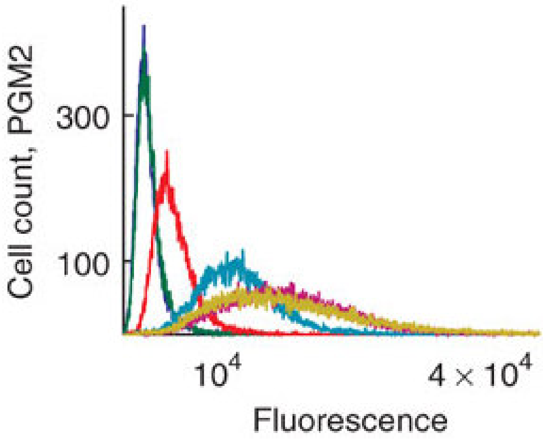


Figure 13: Effect of 3% ethanol on PGM2 expression. Blue, green, red, turquoise, magenta and yellow lines correspond to fluorescence distributions after 0, 30, 60, 90, 120 and 150 minutes from perturbation start, respectively. Taken from Bar-Even et al., 2006.

If the modelling of stochastic gene expression and its inherent assumptions are disregarded, the data can be simply interpreted as an increase in cellular heterogeneity due to environmental changes. Interestingly, the authors found that the stress genes did not follow the noise decrease/abundance increase trend and explained this as a potential lack of strict control of these 'dispensable' genes. Another interpretation of this data could be that a minority of cells experiencing poor mixing conditions, high shear forces and insufficient aeration in culture, specifically expressed their stress genes in a controlled deterministic response.

Implications

A point noted by Golding and colleagues (Golding et al., 2005) regards the use of synthetic promoters to study transcription kinetics. They argue that this minimises the effects of endogenous cellular control circuits, a potential effect previously noted here in the Raser and O'Shea study (Raser and O'Shea, 2004). Golding and colleagues also question the biological relevance of their findings. A recurring theme in the study of gene expression variation is the benefit of synthetic promoters (lactose/IPTG, lambda phage, tetracycline repressor) due to their well characterised responses and the ability to tune their expression (Becskei and Serrano, 2000; Ozbudak et al., 2002; Rosenfeld et al., 2005; Pedraza and van Oudenaarden, 2005). The reporters are then expressed at low levels, or observed in the repressed state, to avoid the rapid saturation of fluorescent signal if reporters are maximally expressed. Therefore, what is being observed could be likened to minimal 'leaky' expression, raising questions about the importance of stochastic mechanisms (if present) to normal gene expression in the cell. These repressors are also linked to systems where stochastic or occasional expression has been proposed to benefit the cell by random sampling of a fluctuating environment (Thattai and van Oudenaarden, 2004) and may reflect the exception, rather than the rule, in bacterial physiology. Tuning of these circuits with synthetic inducers may further amplify artefacts present in the experimental setup. These phenomena may be useful for engineering purposes in synthetic biology, while their implications for understanding natural processes may be limited. Could it be that the synthetic systems used to study gene expression have been unintentionally selected for by researchers because of their highly variable expression dynamics?

If we do assume that stochasticity plays a role in gene expression, we must ask what effect this would have on protein activity, rather than protein concentration. The majority of (non-structural) proteins carry out their functions as enzymes, catalyzing reactions with great efficiency. Once protein levels are at a high enough level, so as to saturate a reaction, small variation in protein level may not have an effect on the reaction rates. This raises a question about the relevence of these fluctuations in maintaining control of normal cellular functions. One example of the potential lack of these effects is a modelling experiment that looked at the ability of oscillators to buffer noise (Vilar et al., 2002). It was suggested that the proteins themselves could effectively attenuate any noise produced at the gene expression level. Another example where noise propagation from transcription may have little or no effect is on signalling modules (Bruggeman et al., 2009), where noise from one level (transcription) may be unable to affect noise at another (protein signalling cascades). However, there does appear to be an enrichment of transcriptional network motifs that are proposed to attenuate noise Kittisopikul and Suel, 2010). Again, this work is susceptible to assumption errors due to the stochastic simulations employed in the study, with a simpler model originally proposed some time earlier (Savageau, 1998).

Another attitude that could be taken is: does it matter if stochasticity is assumed when the models work? They do a better job modelling the data after all. I believe the answer to this question is based on scales, and the level of detail needed for the question posed. This topic is touched upon by Pedraza and Paulsson when they mention the use of 'coarse grained' models and the care that must be taken in the interpretation of results (Pedraza and Paulsson, 2008). If stochastic simulations are used to incorporate an unknown factor into a model they are acceptable. However, if that factor is then assumed to be random, without proper investigation into its cause, the detailed understanding of a system may be limited. Complex systems are known to have emergent properties that can be difficult to comprehend based solely on the sum of their parts. Claiming that the behaviours seen are therefore random, does not do justice to the orchestrated processes taking place inside a cell or population of cells. Assuming stochastic models are employed in a careful manner the question then becomes: what is the level of detail that will satisfy our curiosity, if this is possible at all?

Future directions

The study of stochastic gene expression highlights several important topics in the study of cell biology. Other potential sources of cellular heterogeneity that have non-stochastic drivers may be overlooked by investigators with a stochastic gene expression agenda. Experimental conditions are assumed to consist of perfectly mixed environments containing perfectly isogenic cells with no history, however, each cell can be considered as unique, with a history of its own, however small the difference from its neighbours may be (Jacob, 1977). Sources of microbial cell individuality include ultradian rhythms, epigenetic modifications, ageing linked to cell division, mitochondrial activity and cell specific growth rates (Avery, 2006), factors frequently not included in the modelling or interpretation of data. Isogenic populations and the rates of mutation are another experimental concern (Dekel and Alon, 2005), especially difficult to account for on a single cell level through use of the usual laboratory techniques employed to confirm the fidelity of constructs (Lynch et al., 2008). These are important factors that need to be examined before a potential contribution of stochasticity to gene expression can be confirmed.

Genome wide approaches in cell biology are currently limited to the investigation of mRNA or protein levels across an entire cell population. This limitation is due to the small amount of material that can be extracted from a single cell. Concerns have been raised over the interpretation of results from this hypothetical 'average cell' (Levsky and Singer, 2003). However, if transcription is generally 'bursty' as was shown to be the case in bacteria (Golding et al., 2005), single fixed cells may not represent a genuine picture of how cells behave over time. In this case genome wide studies examining whole populations might be more representative of the general state of the cell, effectively averaging out these temporal fluctuations. Approaches towards investigating stochasticity thus far have looked at the expression of a single gene in a single cell (Elowitz et al., 2002), the expression of several genes in a single cell (Pedraza and van Oudenaarden, 2005), or the expression of many different genes in many cells (but looking at one gene per population) (Newman et al., 2006). To determine the implications of intercellular protein variation the levels of many genes need to be measured in single cells, preferably in real-time. This is clearly not possible at present due to technical limitations. However, progress is being made in the interrogation of single cells through the use of multiplexed qPCR (Diercks et al., 2009) and multiplexed FISH (Levsky et al., 2007), while the incredible goal of a single cell transcriptome has been reached using mRNA-seq (Tang et al., 2009), albeit with a rather large mouse blastomere. Unfortunately, all of these techniques require the extraction of real-time kinetics from a fixed cell analysis, leaving the mathematical interpretations of the results open to questions once more. As real-time visualization of the various products of many genes is technically not feasible at this time, the best surrogate may be to look at the average cell phenotype.

Conclusions

Invoking stochastic mechanisms of gene expression in the search for causes of cellular heterogeneity can result in a circular logic argument of cause and effect. While variation in gene expression has been amply proven, to classify the cause of this variation as stochastic, potentially diminishes our view of a cell's ability to both respond to its environment and control its internal state. To properly measure stochasticity we would need a perfect experiment, with all aspects being initially equal. This is clearly not possible. Reducing the cause of molecular events to a deterministic or stochastic event rapidly becomes a philosophical debate, where determinism, free will and belief systems cloud the topic. It is this author's belief that the study of stochasticity is worthwhile because it questions the underlying causes of cellular heterogeneity. Researchers often refer to biological noise without defining what they really mean by it. As biology is moving towards a systems approach, care must exercised in model choice, so as not to make overreaching interpretations. Cooperation between physicists and biologists has the potential for fruitful results, but assumptions made in mathematical modelling techniques must be realistic from the biological and experimental perspective.

References

Avery, S.V. (2006). Microbial cell individuality and the underlying sources of heterogeneity. Nat. Rev. Microbiol. *4,* 577-587.

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

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

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

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

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H., and Long, R.M. (1998). Localization of ASH1 mRNA particles in living yeast. Mol. Cell *2,* 437-445.

Billinton, N., and Knight, A.W. (2001). Seeing the wood through the trees: a review of techniques for distinguishing green fluorescent protein from endogenous autofluorescence. Anal. Biochem. *291,* 175-197.

Bratsun, D., Volfson, D., Tsimring, L.S., and Hasty, J. (2005). Delay-induced stochastic oscillations in gene regulation. Proc. Natl. Acad. Sci. U. S. A. *102,* 14593-14598.

Bruggeman, F.J., Bluthgen, N., and Westerhoff, H.V. (2009). Noise management by molecular networks. PLoS Comput. Biol. *5,* e1000506.

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

Dekel, E., and Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. Nature *436,* 588-592.

Derouazi, M., Martinet, D., Besuchet Schmutz, N., Flaction, R., Wicht, M., Bertschinger, M., Hacker, D.L., Beckmann, J.S., and Wurm, F.M. (2006). Genetic characterization of CHO production host DG44 and derivative recombinant cell lines. Biochem. Biophys. Res. Commun. *340,* 1069-1077.

Diercks, A., Kostner, H., and Ozinsky, A. (2009). Resolving cell population heterogeneity: real-time PCR for simultaneous multiplexed gene detection in multiple single-cell samples. PLoS One *4,* e6326.

Eldakak, A., Rancati, G., Rubinstein, B., Paul, P., Conaway, V., and Li, R. (2010). Asymmetrically inherited multidrug resistance transporters are recessive determinants in cellular replicative ageing. Nat. Cell Biol. *12,* 799-805.

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

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem *81,* 2340–2361.

Golding, I., and Cox, E.C. (2006). Physical nature of bacterial cytoplasm. Phys. Rev. Lett. *96,* 098102.

Golding, I., and Cox, E.C. (2004). RNA dynamics in live Escherichia coli cells. Proc. Natl. Acad. Sci. U. S. A. *101,* 11310-11315.

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

Horan, B.L. (1994). The Statistical Character of Evolutionary Theory. Philosophy of Science *Vol. 61,* 76-95.

Huffaker, R.G. (1998). Deterministic Modeling without (Unwarranted) Apology. Review of Agricultural Economics *Vol. 20,* 502-512.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425,* 686-691.

Ihmels, J., Levy, R., and Barkai, N. (2004). Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nat. Biotechnol. *22,* 86-92.

Jacob, F. (1977). Evolution and tinkering. Science *196,* 1161-1166.

Kittisopikul, M., and Suel, G.M. (2010). Biological role of noise encoded in a genetic network motif. Proc. Natl. Acad. Sci. U. S. A. 107, 13300-13305.

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

Lee, M.W., Kim, B.J., Choi, H.K., Ryu, M.J., Kim, S.B., Kang, K.M., Cho, E.J., Youn, H.D., Huh, W.K., and Kim, S.T. (2007). Global protein expression profiling of budding yeast in response to DNA damage. Yeast *24,* 145-154.

Levsky, J.M., Shenoy, S.M., Chubb, J.R., Hall, C.B., Capodieci, P., and Singer, R.H. (2007). The spatial order of transcription in mammalian cells. J. Cell. Biochem. *102,* 609-617.

Levsky, J.M., and Singer, R.H. (2003). Gene expression and the myth of the average cell. Trends Cell Biol. *13,* 4-6.

Llopis, J., McCaffery, J.M., Miyawaki, A., Farquhar, M.G., and Tsien, R.Y. (1998). Measurement of cytosolic, mitochondrial, and Golgi pH in single living cells with green fluorescent proteins. Proc. Natl. Acad. Sci. U. S. A. *95,* 6803-6808.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., Dopman, E.B., Dickinson, W.J., Okamoto, K., Kulkarni, S., Hartl, D.L., and Thomas, W.K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. U. S. A. *105,* 9272-9277.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441,* 840-846.

Novick, A., and Weiner, M. (1957). Enzyme Induction as an All-Or-None Phenomenon. Proc. Natl. Acad. Sci. U. S. A. *43,* 553-566.

Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. Nat. Genet. *31,* 69-73.

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

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

Pedraza, J.M., and Paulsson, J. (2008). Effects of molecular memory and bursting on fluctuations in gene expression. Science *319,* 339-343.

Pedraza, J.M., and van Oudenaarden, A. (2005). Noise propagation in gene networks. Science *307,* 1965-1969.

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

Rao, C.V., Wolf, D.M., and Arkin, A.P. (2002). Control, exploitation and tolerance of intracellular noise. Nature *420,* 231-237.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304,* 1811-1814.

Richard, P. (2003). The rhythm of yeast. FEMS Microbiol. Rev. *27,* 547-557.

Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., and Elowitz, M.B. (2005). Gene regulation at the single-cell level. Science *307,* 1962-1965.

Rossi, F.M., Kringstein, A.M., Spicher, A., Guicherit, O.M., and Blau, H.M. (2000). Transcriptional control: rheostat converted to on/off switch. Mol. Cell *6,* 723-728.

Savageau, M.A. (1998). Demand theory of gene regulation. I. Quantitative development of the theory. Genetics 149, 1665-1676.

Sexton, T., Schober, H., Fraser, P., and Gasser, S.M. (2007). Gene regulation through nuclear organization. Nat. Struct. Mol. Biol. *14,* 1049-1055.

Shahrezaei, V., and Swain, P.S. (2008). Analytical distributions for stochastic gene expression. Proc. Natl. Acad. Sci. U. S. A. *105,* 17256-17261.

Snijder, B., Sacher, R., Ramo, P., Damm, E.M., Liberali, P., and Pelkmans, L. (2009). Population context determines cell-to-cell variability in endocytosis and virus infection. Nature *461,* 520-523.

> Note: Interestingly, a commonly used unit of viral activity (multiplicity of infection or MOI) was originally thought to be the result of a stochastic or random Poisson process itself. An exponentially decaying distribution of gene copy numbers were observed in infected cells, leading researchers to assume a random process was at work. It was only after infection was studied in the context of cellular location in a population that the cause was found to be deterministic (Snijder et al., 2009).

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell *9,* 3273-3297.

Steger, D.J., Haswell, E.S., Miller, A.L., Wente, S.R., and O'Shea, E.K. (2003). Regulation of chromatin remodeling by inositol polyphosphates. Science *299,* 114-116.

Sutherland, H., and Bickmore, W.A. (2009). Transcription factories: gene expression in unions? Nat. Rev. Genet. *10,* 457-466.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., and Surani, M.A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6,* 377-382.

Thattai, M., and van Oudenaarden, A. (2004). Stochastic gene expression in fluctuating environments. Genetics *167,* 523-530.
Valegard, K., Murray, J.B., Stonehouse, N.J., van den Worm, S., Stockley, P.G., and Liljas, L. (1997). The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. J. Mol. Biol. *270,* 724-738.

van Hoek, M., and Hogeweg, P. (2007). The effect of stochasticity on the lac operon: an evolutionary perspective. PLoS Comput. Biol. *3,* e111.

Vilar, J.M., Kueh, H.Y., Barkai, N., and Leibler, S. (2002). Mechanisms of noise-resistance in genetic oscillators. Proc. Natl. Acad. Sci. U. S. A. *99,* 5988-5992.

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

Zhang, Q., Andersen, M.E., and Conolly, R.B. (2006). Binary gene induction and protein expression in individual cells. Theor. Biol. Med. Model. *3,* 18.