Lenive *et al. BMC Systems Biology* (2016) 10:81
DOI 10.1186/s12918-016-0324-x

BMC Systems Biology

RESEARCH ARTICLE Open Access

Inferring extrinsic noise from single-cell gene expression data using approximate Bayesian computation

Oleg Lenive[1], Paul D. W. Kirk[2] and Michael P. H. Stumpf[3*]

Abstract

Background: Gene expression is known to be an intrinsically stochastic process which can involve single-digit numbers of mRNA molecules in a cell at any given time. The modelling of such processes calls for the use of exact stochastic simulation methods, most notably the Gillespie algorithm. However, this stochasticity, also termed "intrinsic noise", does not account for all the variability between genetically identical cells growing in a homogeneous environment.
Despite substantial experimental efforts, determining appropriate model parameters continues to be a challenge. Methods based on approximate Bayesian computation can be used to obtain posterior parameter distributions given the observed data. However, such inference procedures require large numbers of simulations of the model and exact stochastic simulation is computationally costly.
In this work we focus on the specific case of trying to infer model parameters describing reaction rates and extrinsic noise on the basis of measurements of molecule numbers in individual cells at a given time point.

Results: To make the problem computationally tractable we develop an exact, model-specific, stochastic simulation algorithm for the commonly used two-state model of gene expression. This algorithm relies on certain assumptions and favourable properties of the model to forgo the simulation of the whole temporal trajectory of protein numbers in the system, instead returning only the number of protein and mRNA molecules present in the system at a specified time point. The computational gain is proportional to the number of protein molecules created in the system and becomes significant for systems involving hundreds or thousands of protein molecules.

Conclusions: We employ this simulation algorithm with approximate Bayesian computation to jointly infer the model's rate and noise parameters from published gene expression data. Our analysis indicates that for most genes the *extrinsic* contributions to noise will be small to moderate but certainly are non-negligible.

Keywords: Stochastic simulation, Gene expression, Extrinsic noise, Approximate Bayesian computation

Background

Experiments have demonstrated the presence of considerable cell-to-cell variability in mRNA and protein numbers [1–5] and slow fluctuations on timescales similar to the cell cycle [6, 7]. Broadly speaking, there are two plausible causes of such variability. One is the inherent stochasticity of biochemical processes which are dependent on small numbers of molecules. The other relates to differences in numbers of protein, mRNA, metabolites and other molecules available for each reaction or process within a cell, as well as any heterogeneity in the physical environment of the cell population. These sources of variability have been dubbed as "intrinsic noise" and "extrinsic noise", respectively.

One of the earliest investigations into the relationship between intrinsic and extrinsic noise employed two copies of a protein with different fluorescent tags, expressed from identical promoters equidistant from the replication origin in *E. coli* [8]. By quantifying fluorescence for a range of expression levels and genetic backgrounds the authors

*Correspondence: m.stumpf@imperial.ac.uk
[3]Imperial College, London, Centre for Integrative Systems Biology and Bioinformatics, SW7 2AZ London, UK
Full list of author information is available at the end of the article

concluded that intrinsic noise decreases monotonically as transcription rate increases while extrinsic noise attains a maximum at intermediate expression levels. Other studies have considered extrinsic noise in the context of a range of cellular processes including the induction of apoptosis [9]; the distribution of mitochondria within cells [10]; and progression through the cell cycle [11]. From a computational perspective, extrinsic variability has been modelled by linking the perturbation of model parameters to perturbation of the model output using a range of methods, including the Unscented Transform [12] the method of moment closure [13], and density estimation [14].

Taniguchi et al. [7] carried out a high-throughput quantitative survey of gene expression in *E. coli*. By analysing images from fluorescent microscopy they obtained discrete counts of protein and mRNA molecules in individual *E. coli* cells. They provided both the measurements of average numbers of protein and mRNA molecules in a given cell, as well as measurements of cell-to-cell variability of molecule numbers. The depth and scale of their study revealed the influence of extrinsic noise on gene expression levels. The authors demonstrated that the measured protein number distributions can be described by Gamma distributions, the parameters of which can be related to the transcription rate and protein burst size [15]. To quantify extrinsic noise they consider the relationship between the means and the Fano factors of the observed protein distributions. They also illustrate how extrinsic noise in protein numbers may be attributed to fluctuations occurring on a timescale much longer than the cell cycle.

Here we aim to describe extrinsic noise at a more detailed, mechanistic, level using a stochastic model of gene expression. A relatively simple mechanistic model of gene expression may represent mRNA production as a zero order reaction with protein being produced from each mRNA via first order reactions. This can be described as the one-state model since the promoter is modelled as being constitutively active (Fig. 1). In the one-state model, mRNA production is represented by a homogeneous Poisson process and the Fano factor of the mRNA distribution at any time point will be one. However, experimental counts of mRNA molecules in single cells indicate that the Fano factor is often considerably higher than one [7].

Such a description calls for quantitative inference of the model's parameters. We achieve this by relying on the data made available by Taniguchi et al. and employing approximate Bayesian computation (ABC) [16, 17]. One difficulty that arises when trying to investigate the extent and effect of extrinsic noise is that it is difficult to separate it from intrinsic noise. To overcome this confounding effect, the parameters of our model come in two varieties. Firstly, reaction rate parameters describe the probability of events occurring per unit of time. These correspond to the reaction rate parameters of a typical stochastic model which accounts for intrinsic noise. Secondly, noise parameters describe the variability in reaction rate parameters caused by the existence of extrinsic noise. In this model, extrinsic noise is represented by a perturbation of the model's rate parameters using a truncated Gaussian distribution. The magnitude of the perturbation of each rate parameter depends on the corresponding noise parameter, which is closely related to the standard deviation of the relevant Gaussian (see "Methods"). This approach allows us to simultaneously infer the rate parameters and the magnitude of extrinsic noise and may be thought of as an application of mixed effect modelling [18] in the context of exact stochastic simulation.

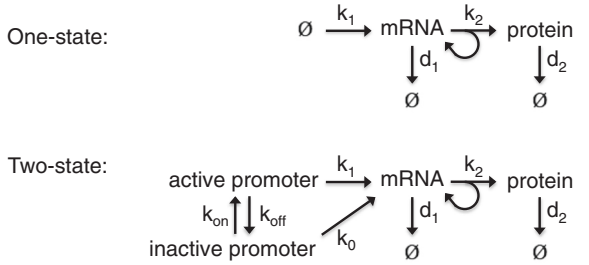

Fig. 1 Schematic representations of the one- and two-state models. In the one-state model (*top*), mRNA is produced at a constant rate (k_1). Protein is produced from mRNA with first order kinetics at a rate k_2. Both mRNA and protein molecules are degraded according to first order kinetics with rates d_1 and d_2 respectively. The two-state model (*bottom*) has the added feature of two promoter states. In the inactive state the promoter produces mRNA at a fraction (k_0) of that of the active state. Switching between the two states corresponds to a telegraph process characterised by two rate constants (k_{on} and k_{off})

Stochastic simulation and ABC inference methods are both computationally costly endeavours. In this particular case, the experimental data corresponds to snapshots of the system at a single time point. The data are made available in the form of summary statistics, measures of central tendency (e.g. mean) and statistical dispersion (e.g. variance).

Thus, a complete temporal trajectory of the system is not necessary to carry out comparisons with the data. This allows us to make the problem computationally tractable. To this end, we develop a model-specific simulation method which takes advantage of the Poissonian relationship between the number of surviving protein molecules produced from a given mRNA molecule and its lifetime, under certain assumptions.

Results and discussion

Posterior distributions of parameters

We begin our analysis by examining the posterior distributions of parameters obtained for each gene using

the ABC Sequential Monte Carlo (ABC-SMC) inference procedure [16] A selection of distributions is shown in Fig. 3 and the Additional files 2, 3, 4 and 5 supplementary figures. The simulated summary statistics converged to within the desired threshold of the experimental measurements for 86 out of 87 genes. The inferred posterior for the one remaining gene converged relatively slowly and we chose to terminate the process after 30 days of CPU time.

Figure 2 shows a contour plot of the distribution of summary statistics and the mRNA degradation rate, obtained from particles in the final ABC-SMC population for a typical gene (*dnaK*).

We begin with a discussion of features of the posterior parameter distributions, that are common to most genes. Next, we examine the relationships between model parameters and summary statistics of the model outputs. Lastly, we carry out a sensitivity analysis on the inferred posteriors to assess the importance of each parameter in setting the overall levels of extrinsic noise.

In the two-state model, the switching of the promoter between active and inactive states is described by a telegraph process that can be parametrised either in terms of the switching reaction rates (k_{on} and k_{off}) or in terms of the on/off bias (k_r) and frequency of switching events (k_f) (Fig. 1). The simulation algorithm takes parameters in the form of k_{on} and k_{off}. However, the effects of k_r and k_f on the observed mRNA distribution may be interpreted more directly and intuitively.

For the majority of genes the k_0 and k_r parameters are relatively small. This appears to be a prerequisite for a high Fano factor of the mRNA distribution and the mean marginal inferred values of these parameters are negatively correlated with Fano factors across all 86 genes as discussed below. A low switching rate combined with a low basal expression rate ensures that there are two distinct mRNA expression levels. This in turn produces a larger variance in measured mRNA counts and results in Fano factor values well above one. Conversely, genes for which mRNA production appears to be more Poissonian were inferred to have basal mRNA production rates close to one, i.e. similar to the active mRNA production rates. In other words, these genes appear to be constitutively active. Here again, we point out that the two-state promoter model provides a convenient abstraction and a hypothesis for explaining the super-Poissonian variance in mRNA copy number [5, 19]. However, based on these observations it is difficult to determine whether a model with more states or some other more elaborate regulatory model, would not be more appropriate. Our attempts

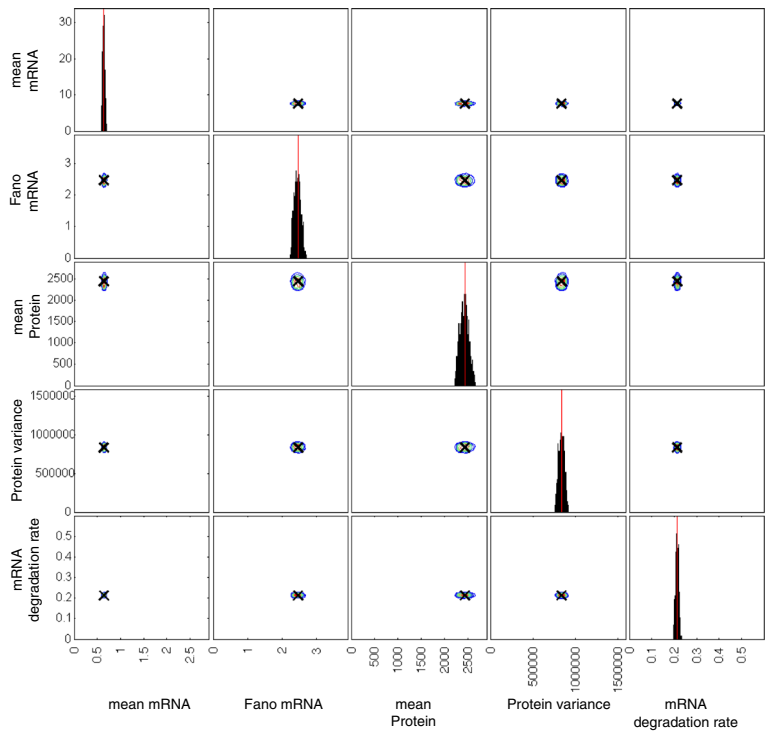

Fig. 2 Posterior distribution of summary statistics and the mRNA degradation rate for the gene *dnaK*. Contour plots indicating the density of points with the corresponding summary statistic for each particle in the final population. The summary statistics for each particle are calculated from 1000 simulation runs. The posterior distribution consists of 1000 particles

at carrying out the inference procedure with a one-state model indicate that extrinsic noise alone does not explain the observed mRNA distributions without also producing unacceptably high variability in protein numbers.

Our initial inference attempts used only the summary statistics from the data. We observed that the production and degradation rate parameters for mRNA (k_1 and d_1) and protein (k_2 and d_2) tended to be positively correlated in the posterior parameter distributions of many genes. This is due to limited identifiability of model parameters since different combinations of rates may produce similar steady state expression levels. We included the mRNA degradation rate in the inference procedure with the aim of overcoming the problem of unidentifiable parameters. However, this did not alleviate the problem entirely and there is still considerable uncertainty, or sloppiness, in the posterior with regard to some directions in parameter space. While this does make it difficult to pick precise parameter values it also illustrates how using ABC provides us with a way of measuring the model's sensitivity to changes in parameters. Our approach provides an indication of the possible range of extrinsic noise values that can account for the observed variability in mRNA and protein numbers (Fig. 3).

Although the posterior summary statistics (and mRNA degradation rate) are reasonably well constrained and distinct for each gene, the distributions of model parameters can still be relatively broad (Fig. 3). There are a number of reasons for this. Firstly, changes in parameters associated with active transcription and translation, as well as degradation rates, are more easily inferred than parameters describing switching between promoter states, basal transcription or extrinsic noise. In particular, when the production and degradation rates for the same species are subjected to different extrinsic noise parameters, the inference procedure struggles to resolve between the different source of extrinsic noise. This explains the correlation between the means of inferred extrinsic noise parameters (Fig. 4). Such correlations between extrinsic noise parameters are not observed in the posterior of each gene or when taking the single particle with the highest weight from the final population of each gene as in Fig. 5.

A comparison of Figs. 4 and 5 suggests that a certain level of extrinsic noise is expected for all genes. However, the extrinsic noise may affect various combinations of rate parameters and it may not be possible to discern if, for example, the production rate or the degradation rate is more affected by extrinsic variability. While our inference procedure does not indicate a distinctive lower boundary for the amount of extrinsic noise affecting each reaction rate, there is usually an upper limit to the inferred noise

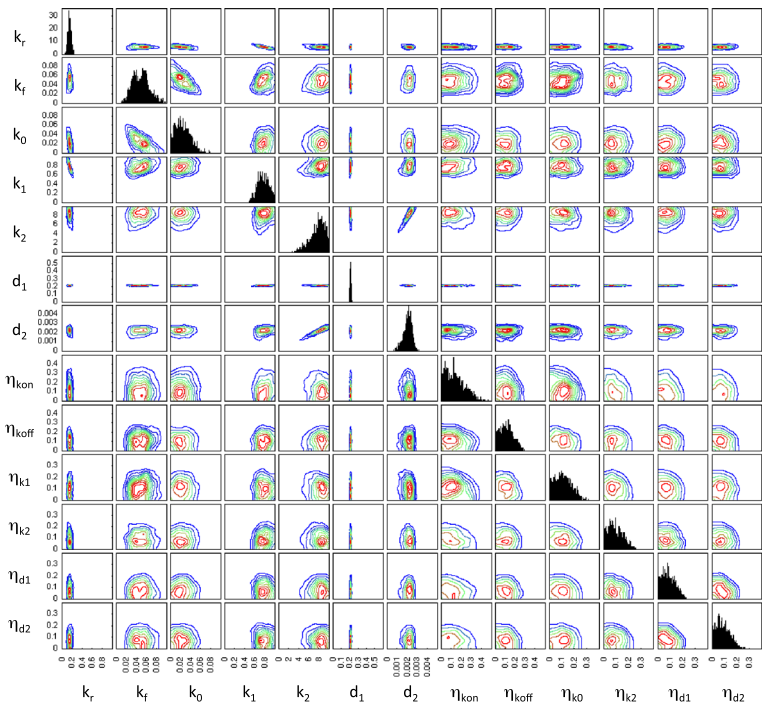

Fig. 3 Posterior distribution of model parameters for the gene *dnaK*. Contour plots indicating the density of points with the corresponding parameter values for each particle in the final population. The posterior distribution consists of 1000 particles

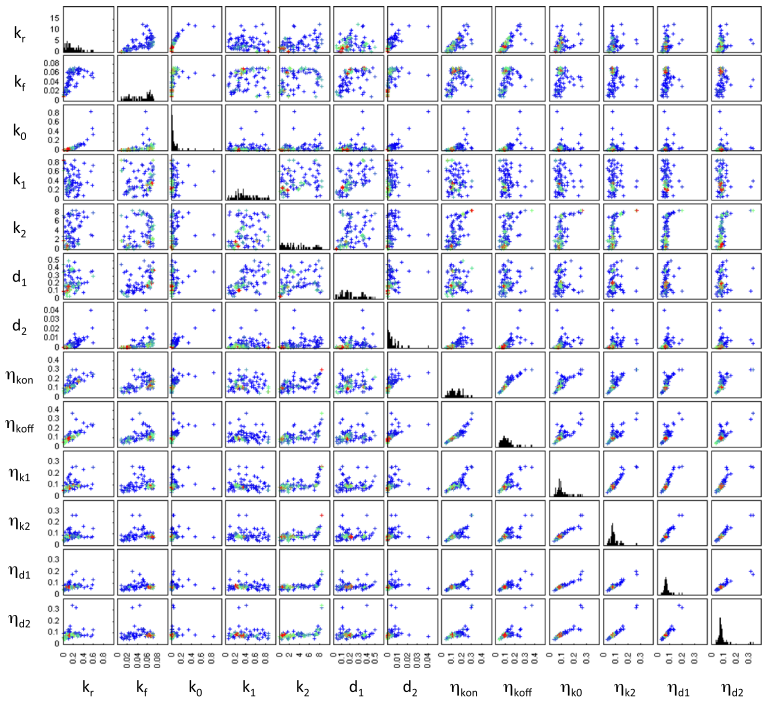

Fig. 4 Relationships between means of the marginal parameter posteriors. Scatter plots of the means of the marginal distributions of parameter posteriors are shown for all pairs of parameters. Each point corresponds to a gene. Warmer hues are used to indicate a higher density of data points

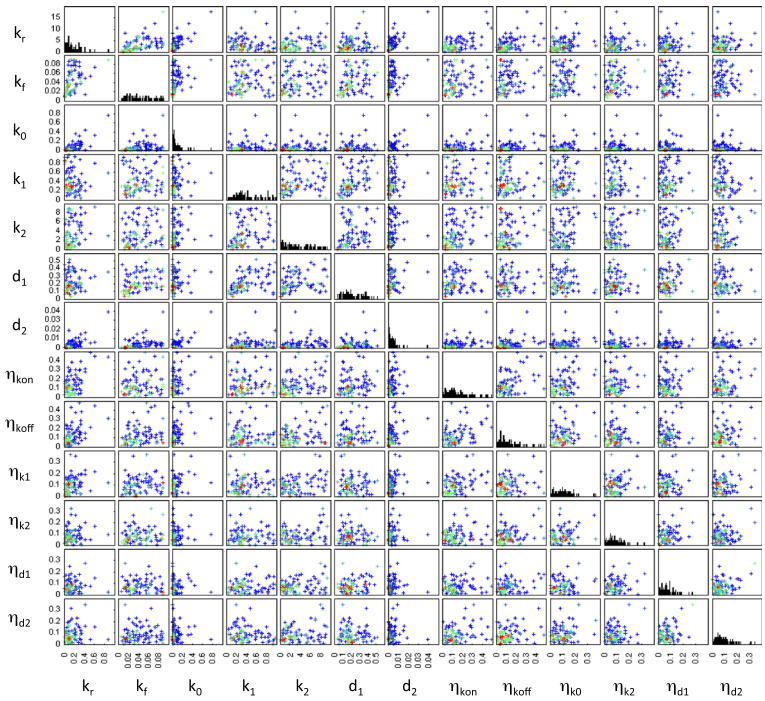

Fig. 5 Relationships between the heaviest particles. Scatter plots of the particles with the highest weight in the final ABC-SMC population, shown for all pairs of parameters. Each point corresponds to one particle from the inferred posterior of one gene. Warmer hues are used to indicate a higher density of data points

parameters ranges. The extrinsic noise parameters for most genes are below 0.2 in the units set here (Fig. 5); however, for some genes, $\eta_{k_{\text{on}}}$ and $\eta_{k_{\text{off}}}$ have relatively broad posterior marginal distributions.

To better understand the relationship between model parameters and observed patterns of gene expression, we look for correlations between means and variances of the inferred marginal parameters of each gene and the summary statistics used in the inference procedure (Fig. 6). As expected, the correlation between the measured mRNA degradation rate, calculated form mRNA lifetime, and the inferred mRNA degradation rate parameter of the model, is close to one.

The promoter switching rate parameters, k_{on} and k_{off}, display positive and negative correlation with the mean mRNA number, respectively (as may be expected). They have the opposite relationship with the Fano factor associated with the mRNA distribution. This is consistent with the idea that distinct levels of transcription are required to account for the observed mRNA Fano factors. The corresponding extrinsic noise parameters $\eta_{k_{\text{on}}}$ and $\eta_{k_{\text{off}}}$ are positively correlated with mRNA abundance. However, the means and variances of the marginal distributions of these parameters are negatively correlated with the Fano factor of the mRNA distribution. This indicates that when promoter switching is affected by higher extrinsic noise, the mRNA distribution becomes more Poissonian as the effect of the two distinct promoter states is averaged out.

Curiously, the mean and variance of the protein degradation rate (d_2) are positively correlated with mean mRNA number and negatively correlated with the mRNA Fano factor. Unlike the translation rate (k_2), it shows no significant correlation with the mean or variance of the protein number.

Parameter sensitivity

There are two complementary approaches to investigating the sensitivity of a modelled system to its parameters or inputs [20]. One approach is to consider a single point in parameter space and study how the model responds to infinitesimal changes in parameters. This local approach usually involves calculating the partial derivatives of the model output with respect to the parameters of interest. Alternatively, one may consider how the model behaviour varies within a region of parameter space by sampling parameters and observing model behaviour. Regardless of the method used, different linear combinations of parameters will affect the model output to varying degrees [21]. Gutenkunst et al. [22] coined the terms "stiff" and "sloppy" to describe these differences. They defined a Hessian matrix,

$$H^{\chi^2}_{i,j} \equiv \frac{d^2\chi^2}{d\log\theta_i d\log\theta_j},$$

where χ^2 provides a measure of model behaviour, such as the average squared change in the species time course.

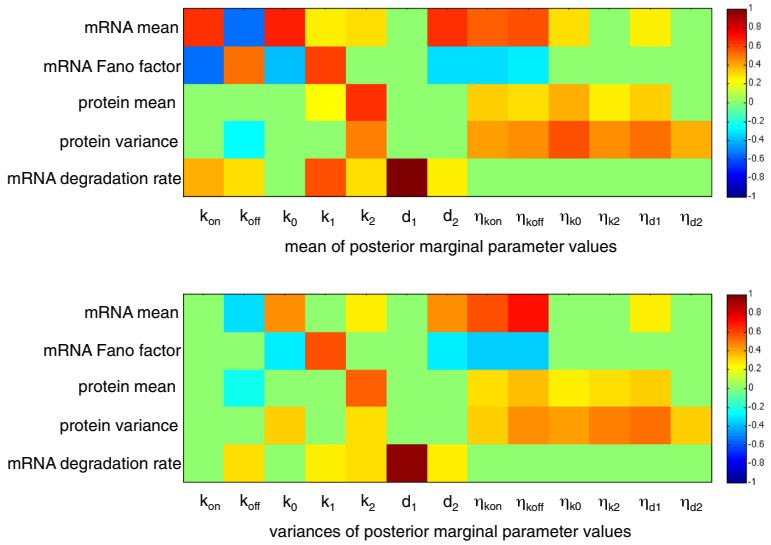

Fig. 6 Heat maps of correlation coefficients between parameters and summary statistics. Heat maps are of the correlation coefficients calculated between experimentally obtained summary statistics and the mean (*top*) or the variance (*bottom*) of the marginal posterior for each model parameter. Correlation coefficients for which the associated *p*-values are greater than 0.05, after correcting for multiple testing using the Benjamini-Hochberg method [43], are treated as zero for plotting purposes

By considering the eigenvalues of this Hessian, λ_i, the authors were able to quantify the (local) responsiveness of the system to a given change in parameters. Conceptually, moving along a stiff direction in parameter space causes a large change in model behaviour; conversely moving along a sloppy direction results in comparatively little effect on the output of the system.

Secrier et al. [23] later demonstrated how these ideas can be applied to the analysis of posterior distributions obtained by ABC methods [24]. Principal component analysis (PCA) may be used to approximate the log posterior density using a multivariate normal (MVN) distribution. They showed that the eigenvalues of the covariance matrix, s_i, of this MVN distribution are related to the eigenvalues of the Hessian as $\lambda_i = 1/s_i$.

To assess the the stiffness/sloppiness of the inferred parameters we carry out PCA of the covariance matrices of log posterior distributions for each gene. In interpreting the results of the PCA we assume that the posterior distribution is, in practice, unimodal. The principal components (eigenvectors), v, and the corresponding loadings (eigenvalues), s, provided by the PCA are then used to obtain the eigen-parameters, q, as

$$q_i = s_i v_i.$$

We calculate the projections of each parameter, θ_i, onto each eigen-parameter, q_j, as

$$c_{ij} = \theta_i \cdot q_j.$$

As a measure of the overall sloppiness of each parameter, l, we use the sum of the contributions of each parameter to the eigen-parameters, $l_i = \sum_j c_{ij}$. This can also be thought of as the sum of the projections of each principal component onto the parameter, weighted by the fraction of total variance explained by each of the principal components.

Having obtained a measure of the sloppiness of each parameter, for each gene, we carry out hierarchical clustering [25] of genes and parameters using a Euclidean distance metric for both (Fig. 7).

The majority of genes show a similar pattern of parameter stiffness/sloppiness. The most distinctive and the second most distinctive clusters consist of just two genes each, *yiiU* with *aceE* and *cspE* with *map*, respectively. These four genes are distinguished by unusually sloppy promoter activity ratio, k_r, and promoter switching frequency, k_f, parameters. The pair *yiiU* and *aceE* display a high ratio of protein variance to protein mean (Fano factor) and are stiff with regard to the protein degradation rate noise parameter η_{d_1}. *cspE* also has a high

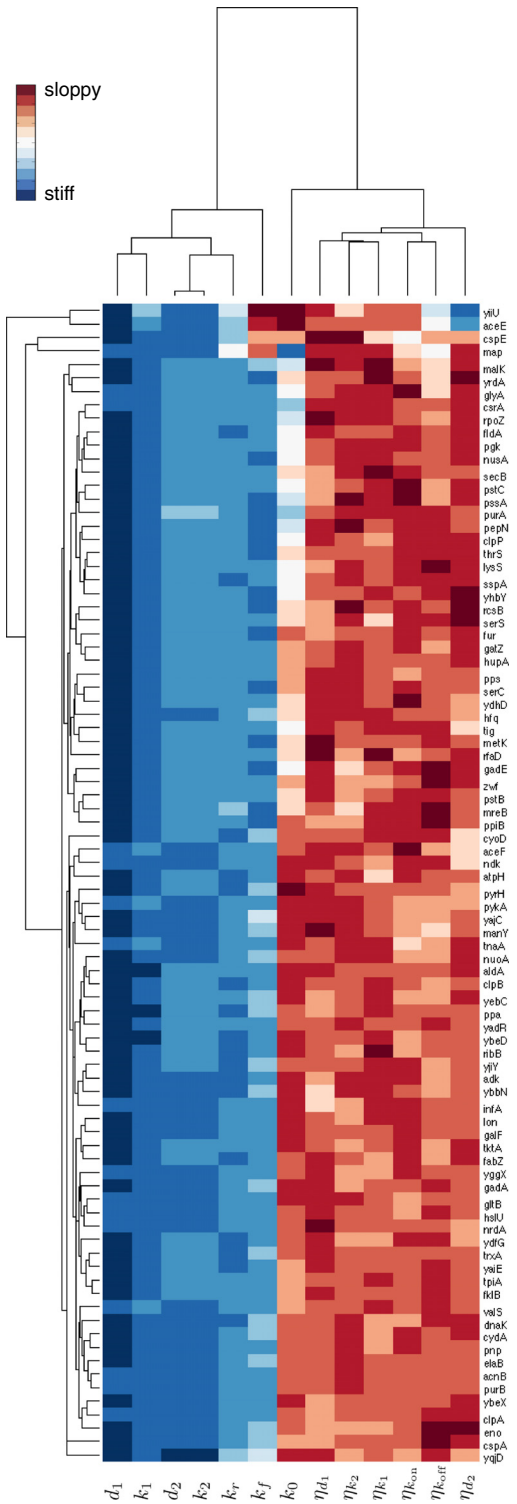

Fig. 7 Clustering of genes and inferred posteriors according to parameter sloppiness. The clustergram shows a heat map of parameter (*columns*) sloppiness for each gene (*rows*). Warmer hues indicate more sloppy parameters. Dendograms above and to the left of the heat map display the hierarchical tree obtained when clustering either the model parameters or the genes using a Euclidean distance metric

Fano factor of the protein distribution while *map* has an unusually low mRNA Fano factor. What these four genes appear to have in common is that the variability in their protein numbers is difficult to explain based solely on the mRNA variability. Thus, a higher level of extrinsic noise is inferred to account for the observed variability. Since these genes comprise a small minority, it may be that their expression is subject to regulatory mechanisms that are not well approximated by the two state model. The remaining majority of genes are broadly divided into two similar groups which differ mostly in the sloppiness of k_0.

The noise and rate parameters segregate into two clusters with the noise parameters generally being sloppier than the rate parameters (Fig. 7). The least sloppy parameter is the mRNA degradation rate (d_1). This is not surprising since it was used, together with the molecule number summary statistics, to infer the posterior distribution. Of the rate parameters, the basal transcription rate (k_0) is the sloppiest and often approaches the noise parameters in its sloppiness. Since this parameter is defined as a fraction of the active transcription rate (k_1), its relative sloppiness should not be equated to a lack of importance. For most genes the marginal posterior of k_0 is largely constrained to the lower half of its prior distribution, $U(0, 1)$. The only exception being the gene *map* for which the measured mRNA Fano factor was close to one and the marginal posterior of k_0 is in the top half of the prior range. The mean of the marginal posterior of k_0 is negatively correlated with the mRNA Fano factor across all genes (Fig. 6). The two other parameters that influence the mRNA Fano factor, k_r and k_f, are the next sloppiest rate parameters.

Conclusions

Cell-to-cell variability in genetically homogeneous populations of cells is a ubiquitous phenomenon [26–28]. Attempts to quantify it are complicated by the difficulty of assigning it to a single cellular process or any one experimentally measurable variable. It can also be difficult, for example, to distinguish between the intrinsic stochasticity of biochemical processes in the short term and longer term variations which may have been inherited from previous cell generations.

By including a representation of extrinsic noise in our model of gene expression we infer the extent to which the rates of biochemical processes can vary between cells while still producing the experimentally measured mRNA and protein variability. We demonstrate the usefulness of an efficient method for exact stochastic simulation of the two-state model of gene expression. The two-state model is necessary to explain the experimentally measured mRNA variation (Fano factor), and is capable of describing the majority of the observed data. The corresponding single-state model, with constant promoter activity and extrinsic noise, does not produce mRNA Fano factors as high as those measured experimentally without leading to unacceptably high variability in the protein numbers. We show that the amount of extrinsic noise affecting most genes appears to be limited, but non-negligible.

The exact simulation method described here occupies a niche between those cases when only samples from the steady state mRNA distribution of the two-state model [3, 29, 30] are required, and cases when an approximation to the protein distribution [15, 31] is sufficient. The computational advantages of the simulation method described here are limited to specific conditions, such as, low numbers of mRNA molecules and higher numbers of protein molecules. The most limiting factor of this simulation method is that it is not applicable to models in which the protein products affect upstream processes such as promoter activity, transcription or translation. The addition of such interactions would mean that the assumptions used in deriving the Poissonian relationship between the number of surviving protein molecules produced form a given mRNA molecule and mRNA's lifetime would no longer be satisfied. Perhaps an approximate algorithm could be developed on the basis of Algorithm (1) to handle such situations. Alternatively, the tau-leaping algorithm [32], or moment expansion [33, 34], may be more appropriate for models involving these kinds of feedback interactions. Algorithm (1) could, however, be naturally extended to models involving regulatory interactions between non-coding RNAs as the simulation of that part of the model is equivalent to Gillespie's exact algorithm. Although here we use summary statistics of mRNA and protein number measurements, the simulation method is also applicable to cases where a direct comparison between sample distributions, for example using the Hellinger distance, is required.

Here we have worked under the assumption that experimental measurement error associated with individual mRNA or protein counts obtained by fluorescence microscopy are small relative to the combined effects of extrinsic and intrinsic noise. We deem this to be justifiable given the experimental method used by Tanaguchi et al. [7] and the results presented in their publication. More generally, such measurement errors would inflate estimates of the variances in molecule numbers and may skew the inferred extrinsic noise parameters. Other studies, which look directly at the interplay between intrinsic and extrinsic noise in single cells [35] — using time-resolved proteomics data — do also bear this out.

The inferred extrinsic noise parameters will also include the effects of regulatory mechanisms that are not well described by the two-state model. In this sense, our definition of noise becomes blurred with our ignorance about

the regulatory interactions involved in the expression of each gene. Nonetheless, the biochemical mechanisms governing gene expression in a given species are shared between many genes. This is in agreement with our observation that, for most genes, inferred model parameters show similar patterns of sloppiness. If we are able to refine our understanding of the shared aspects of gene expression, we may be able to improve our understanding of both the nature of the noise affecting it, and the regulatory mechanisms controlling it. In practice this may mean finding a mechanistic explanation for the two-state model or further refining it to achieve a better agreement between simulations and experimental results.

The in silico approach used here not only relied on, but was inspired by the experimental work of Tanaguchi et al. [7]. As the resolution of high throughput experimental techniques and the quantity of data they generate continues to increase, more complete observations of cellular processes may begin to yield data amenable to statistical analysis and inference of extrinsic noise. These may in turn require other modelling, computational and theoretical approaches which would not rely on the assumptions and simplifications that we make in this work [36].

Methods

Modelling gene expression

A simple model of gene expression may represent the processes of transcription and translation using mass-action kinetics to describe production and degradation of various species as pseudo-first order reactions. Such a model may be simulated stochastically to take into account the intrinsic variability of processes involving low numbers of molecules. In the simplest version of this model, mRNA is produced from the promoter at a constant rate. However, such Poissonian mRNA production is often not sufficient to account for the variability in mRNA numbers measured experimentally in both prokaryotic and eukaryotic cells. In addition to this, for many genes, transcription appears to occur in bursts rather than at a constant rate. These characteristics of gene expression have been observed in organisms as diverse as bacteria [7], yeast [4], amoeba [2] and mammals [3]. One model of gene expression that takes this into account is the, so called, two-state model.

The two-state promoter model

In the two-state model of gene expression, a gene's promoter is represented as either active or inactive [5, 19]. Here we use a variant of the two-state model with the inactive state corresponding to a lower transcription rate rather than no transcription at all. For each state of the promoter, transcription events at that promoter are represented by a Poisson process with rate parameter corresponding to the transcription rate. Biochemical processes such as transcription factor binding or reorganisation of chromatin structure may account for the existence of several distinct levels of promoter activity. However, which factors play a dominant role in the apparent switching, remains an unanswered question.

The Gillespie algorithm [37] may be used to simulate all the reactions represented by this model and obtain a complete trajectory of the system through time. However, in this case we are only interested in the number of molecules present at the time of measurement. We use a model-specific stochastic algorithm (Algorithm 1) which allows us to reduce the number of computational steps required to obtain a single realisation from the model.

The following reactions, represented using mass-action kinetics, comprise the two-state model:

$$
\begin{aligned}
\texttt{inactive-promoter} &\xrightarrow{k_{\text{on}}} \texttt{active-promoter}\\
\texttt{active-promoter} &\xrightarrow{k_{\text{off}}} \texttt{inactive-promoter}\\
\texttt{inactive-promoter} &\xrightarrow{k_0} \texttt{inactive-promoter} + \texttt{mRNA}\\
\texttt{active-promoter} &\xrightarrow{k_1} \texttt{active-promoter} + \texttt{mRNA}\\
\texttt{mRNA} &\xrightarrow{k_2} \texttt{mRNA} + \texttt{Protein}\\
\texttt{mRNA} &\xrightarrow{d_1} \emptyset\\
\texttt{Protein} &\xrightarrow{d_2} \emptyset
\end{aligned}
$$

The propensity functions (hazards) for each of the above reactions are listed below:

$$
\begin{aligned}
h_0 &= k_{\text{on}}[\texttt{inactive-promoter}]\\
h_1 &= k_{\text{off}}[\texttt{active-promoter}]\\
h_2 &= k_0[\texttt{inactive-promoter}]\\
h_3 &= k_1[\texttt{active-promoter}]\\
h_4 &= k_2[\texttt{mRNA}]\\
h_5 &= d_1[\texttt{mRNA}]\\
h_6 &= d_2[\texttt{Protein}]
\end{aligned}
$$

Here the square brackets refer to the number of molecules of a species rather than its concentration.

The model presented here relies on a number of assumptions about the process of gene expression. Firstly, that the production of mRNA and protein can be described sufficiently well by pseudo-first order reactions. Secondly, that degradation of mRNA and protein can be described as an exponential decay. In a bacterial cell, mRNA molecules are degraded enzymatically and typically have a half-life on the scale of several minutes. The

half-life of protein molecules usually exceeds the time required for cell growth and division during the exponential growth phase. Thus, dilution due to partitioning of protein molecules between daughter cells tends to be the dominant factor in decreasing the number of protein molecules. Here we do not build an explicit model of cell division, instead the decrease in protein numbers is approximated by an exponential decay. Finally, it is assumed that there is no feedback mechanism by which the number of mRNA or protein molecules produced by the gene affects its promoter switching, transcription or translation rates.

Representing extrinsic noise

We model extrinsic noise by perturbing the reaction rate parameters, using a Gaussian kernel, before each simulation of the model [35, 38]. The effect of extrinsic noise on each reaction is assumed to be independent. The reaction rates associated with a particular gene are termed nominal parameters (θ_n).

$$\theta_n = [\, k_{\text{on}}, k_{\text{off}}, k_0, k_1, k_2, d_1, d_2]$$

The values determining the magnitude of the perturbation are termed the noise parameters (η).

$$\eta = [\, \eta_{k_{\text{on}}}, \eta_{k_{\text{off}}}, \eta_{k_0}, \eta_{k_2}, \eta_{d_1}, \eta_{d_2}]$$

Together they comprise the full parameter set for the model $\theta = [\, \theta_n, \eta]$.

In the case of the two-state model of a single gene, each θ_n has a corresponding extrinsic noise parameter with the exception that the basal transcription rate (k_0') is defined as a fraction of the active transcription rate (k_1') so the two reaction rates are subject to the same perturbation (η_{k_0}) before each simulation. This is motivated by the idea that extrinsic factors affecting the transcription rate do not depend on the state of the promoter. The parameters used to generate a single realisation from the two-state model are obtained by sampling from $f(\mu, \sigma)$. Where f is a truncated normal distribution, restricted to non-negative values by rejection sampling, with μ and σ being the mean and standard deviation of the corresponding normal distribution.

$$\begin{aligned}
k_{\text{on}}' &\sim f\left(k_{\text{on}}, k_{\text{on}}\eta_{k_{\text{on}}}\right) \\
k_{\text{off}}' &\sim f\left(k_{\text{off}}, k_{\text{off}}\eta_{k_{\text{off}}}\right) \\
k_1' &\sim f\left(k_1, k_1\eta_{k_1}\right) \\
k_0' &= k_0 k_1' \\
k_2' &\sim f\left(k_2, k_2\eta_{k_2}\right) \\
d_1' &\sim f\left(d_1, k_1\eta_{d_1}\right) \\
d_2' &\sim f\left(d_2, k_2\eta_{d_2}\right)
\end{aligned}$$

The final time point of each simulation represents the number of mRNA and protein molecules in a single cell at the time of measurement.

Simulation procedure

In order to reduce the computational cost of each simulation, rather than using Gillespie's direct method to simulate the entire trajectory of mRNA and protein numbers, we employed Algorithm 1 to obtain samples of the numbers of mRNA and protein molecules at the time of measurement (t_m). First, a realisation of the telegraph process is used to obtain the birth and decay times of mRNA molecules. These are then used to sample the number of protein molecules that were produced from each mRNA molecule and survived until t_m. This procedure makes use of the Poisson relationship between the life time of an individual mRNA molecule and the number of surviving protein molecules that were produced from it. This relationship is derived in Additional file 1 and its use is illustrated in Fig. 8. The final result is the number of both mRNA (M) and protein (P) molecules present in the system at t_m.

Use of experimental data

Using an automated fluorescent imaging assay, Taniguchi et al. [7] were able to quantify the abundances of 1018 proteins from a yellow fluorescent protein fusion library. We focus on a subset of 87 genes from the published data set from [7]. These are all the genes for which, in addition to protein numbers, the experimental data include both fluorescence *in situ* hybridization measurements [39] of mRNA numbers and mRNA lifetimes measurements obtained using RNAseq [40]. We note that these genes are not a random sample from the set of all genes and exhibit higher than average expression levels.

To identify model parameters for which the two-state model, with extrinsic noise, is able to reproduce the experimental measurements, we carry out Bayesian inference using an ABC sequential Monte Carlo (SMC) algorithm that compares summary statistics from simulated and experimental data [41]. Specifically we used the following summary statistics: (1) the mean numbers of mRNA molecules; (2) the Fano factors of mRNA molecule distributions; (3) the mean numbers of protein molecules; (4) the variances of protein molecule numbers; and (5) mRNA lifetimes converted to expontial decay rate parameters. The distributions of these summary statistics are shown in Fig. 9. We assume that the summary statistics correspond to steady state expression levels for each gene. While there is no guarantee that this is the case for every gene, the majority of genes are unlikely to be undergoing major changes in their expression level given that the cells are in a relatively constant environment.

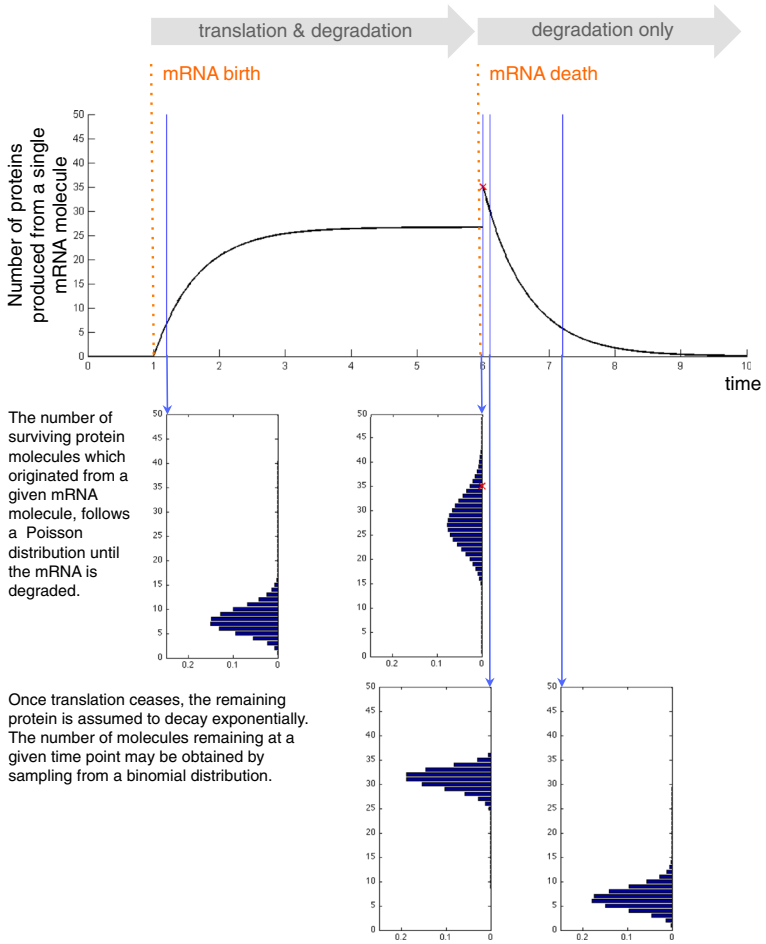

Fig. 8 Illustration of the principle behind Algorithm 1. An illustration of how the birth and death times of an mRNA molecule are used to obtain the number of proteins that were produced from it and then survived until the time at which mRNA and protein numbers were measured. According to the two-state model used here, the number of protein molecules that were translated using a given mRNA template and have not yet been degraded can be found by sampling from the corresponding Poisson distribution with a parameter which depends on the lifetime of the mRNA template. If the mRNA is degraded before the measurement time point, the remaining protein molecules are assumed to decay exponentially. Thus the number of protein molecules can be obtained by first sampling the number present at the point of mRNA decay and then sampling from the corresponding binomial distribution to determine the number of surviving molecules at the measurement time point

Taniguchi et al. [7] used images of about a thousand cells to obtain estimates of mean mRNA numbers, mRNA Fano factors, mean protein numbers and protein number variances. For this reason, we use 10^3 simulation runs when calculating summary statistics. The experimental measurements of mRNA lifetimes are compared directly to the mRNA degradation rate parameter (d_1) in the model by assuming that lifetimes correspond to the inverse of the decay rate.

Inference procedure

We use an ABC-SMC algorithm to infer plausible parameter sets for the two-state model based on the experimental data. The inference procedure is similar to that employed by [24, 41, 42], as described in Algorithm 2.

For the distance metric, d, we take the Euclidean distance between the logarithms of each type of experimental measurement (D_i) and the corresponding simulation results (x_i):

$$d(D,x) = \sqrt{\sum_{i=1}^{i=5} \left(\log D_i - \log x_i\right)^2}$$

$$D = \left[\mu_{mRNA}, \frac{\sigma^2_{mRNA}}{\mu_{mRNA}}, \mu_{prot}, \sigma^2_{prot}, \tau^{-1}_{mRNA}\right]$$

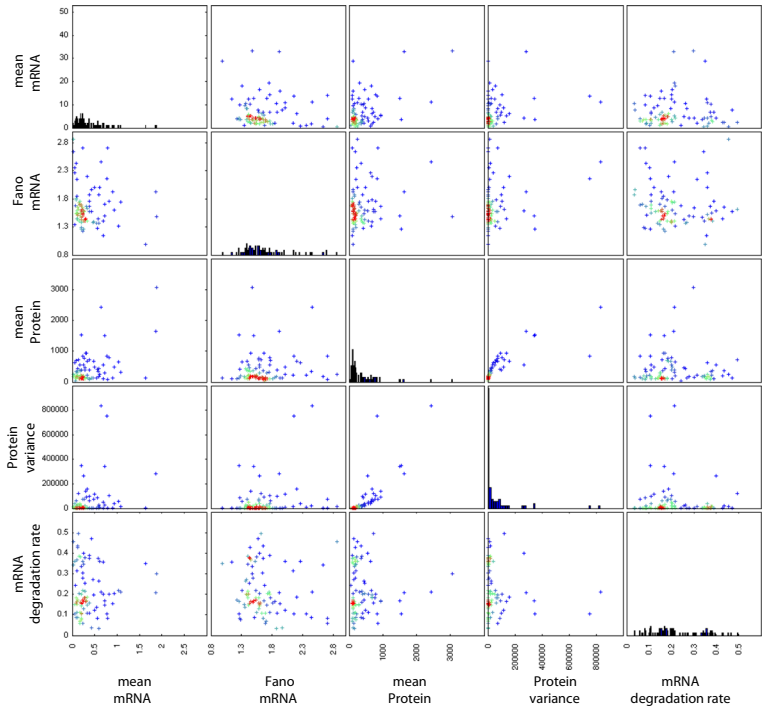

Fig. 9 Experimentally measured summary statistics. Each point on the scatter plots is an estimate of the corresponding summary statistic or mRNA degradation rate from experimental measurements. These data are taken from [7]. The mRNA degradation rates were taken to be the inverse of the mRNA lifetimes

Where μ_{mRNA} is the mean number of mRNA molecules; $\sigma^2_{mRNA}/\mu_{mRNA}$ is the Fano factor of the mRNA distribution; μ_{prot} is the mean number of protein molecules; σ^2_{prot} is the variance of the protein distribution and τ^{-1}_{mRNA} gives the exponential decay rate constant for mRNA degradation based on the measured mRNA lifetime (τ_{mRNA}).

$$x = \left[\mu_M, \frac{\sigma^2_M}{\mu_M}, \mu_P, \sigma^2_P, d_1\right]$$

Where μ_M is the mean number of mRNA molecules; σ^2_M/μ_M is the Fano factor of the mRNA distribution; μ_P is the mean number of protein molecules; σ^2_P is the variance of the protein distribution and d_1 corresponds to the nominal mRNA degradation rate. The first sampled population of particles (population zero in Algorithm 2), provides a benchmark for the choice of ϵ values in the next population. Since we have no knowledge of the distribution of distances until a set of particles is sampled, all particles are accepted in the first population. For subsequent populations, ϵ values are chosen such that the probability of acceptance with the new ϵ value is equal to q_t. The vector q is chosen prior to the simulation. This allows for larger decreases in ϵ in the first few populations while keeping the actual epsilon values used, a function of the distances (g) in the previous population. New populations are sampled until the final epsilon value is reached $\epsilon_f = 0.1$. To obtain θ^* from θ we use a uniform perturbation kernel:

$$\theta^* \sim U(\theta - \mu_{t-1}, \theta + \mu_{t-1})$$

where μ_{t-1} is the vector of standard deviations of each parameter in the previous population.

Parameter prior

The telegraph process may be parametrized in terms of the ratio of probabilities of switching events (k_r) and the overall frequency with which events occur (k_f):

$$k_r = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}}$$

Algorithm 1 Simulation of the two-state model

Inputs: θ_n, η, t_m

Outputs: M, P

Obtain perturbed parameters using the nominal (θ_n) and noise (η) parameters.

Stage one: simulate mRNA production subject to an underlying telegraph process.

$S \leftarrow 1$ with probability $k'_{\text{on}}/(k'_{\text{on}} + k'_{\text{off}})$), otherwise $S \leftarrow -1$ ▷ Select the initial state of the telegraph process.

$t \leftarrow 0$ ▷ Initialise simulation time.

$M_b \leftarrow 0$ ▷ Initialise the number of mRNA molecules produced.

$i \leftarrow 1$ ▷ Initialise index of mRNA molecules.

while $t < t_m$ **do**

  if S = -1 **then**

    $k_S \leftarrow k'_{\text{on}}$

    $k_m \leftarrow k'_0$

  else

    $k_S \leftarrow k'_{\text{off}}$

    $k_m \leftarrow k'_1$

  end if

  $\tau \sim Exp(k_S)$ ▷ Sample the time until the next switching event.

  if $t + \tau > t_m$ **then** ▷ Ensure that $t + \tau$ does not exceed the final time point.

    $\tau \leftarrow t_m - t$

  end if

  $M_\tau \sim Poisson(\tau k_m)$ ▷ Sample the number of mRNA molecules produced.

  $M_b \leftarrow M_b + M_\tau$

  while $i \leq M_b$ **do**

    $u_i \sim Uniform(t, t + \tau)$ ▷ Sample birth times for each mRNA.

    $i \leftarrow i + 1$

  end while

  $t \leftarrow t + \tau$

  $S \leftarrow -S$

end while

Stage two: simulate mRNA degradation; protein production and degradation.

$M \leftarrow 0$ ▷ Initialise the number of mRNA molecules at t_m.

$P \leftarrow 0$ ▷ Initialise the number of protein molecules at t_m.

$i \leftarrow 1$

while $i \leq M_b$ **do** ▷ For each mRNA molecule that was produced:

  $v \sim Exp(d'_1)$ ▷ Sample the time until mRNA decay.

  $T_l \leftarrow min(u_i + v; t_m) - u_i$ ▷ Calculate mRNA lifetime.

  $P_l \sim Poisson\left(\frac{k'_2}{d'_2}(1 - e^{-d'_2 T_l})\right)$ ▷ Sample the number of surviving proteins at time point $u_i + v$.

  $T_d \leftarrow t_m - min(u_i + v; t_m)$ ▷ Time since mRNA decay.

  if $T_d = 0$ **then**

    $M \leftarrow M + 1$ ▷ mRNA survived until t_m.

    $P \leftarrow P + P_l$

  else

    $P_s \sim Binomial(P_l, e^{-d'_2 T_d})$ ▷ Sample the number of surviving proteins at time t_m.

    $P \leftarrow P + P_s$

  end if

  $i \leftarrow i + 1$

end while

Algorithm 2 ABC-SMC with summary statistics

Inputs: π, N, ϵ_f
Outputs: Set of populations of N accepted particles

$i \leftarrow 1$
$t \leftarrow 0$
$q \leftarrow [0.01, 0.05, 0.25, 0.75, ..., 0.75]$
Initialise ϵ vector.
while $\epsilon > \epsilon_f$ **do**
  if $t = 0$ **then**
    while $i \leq N$ **do**
      Sample a new θ from π.
      Simulate from the model 10^3 times according to Algorithm 1.
      Calculate summary statistics, x, from the simulation outputs.
      if $d(D, x) < \epsilon$ **then**
        Accept particle.
        $\omega^{(i,t)} \leftarrow 1$
        $i \leftarrow i + 1$
      end if
    end while
  else
    while $i \leq N$ **do**
      Sample θ from $\{\theta^{(j,t-1)}\}_{1 \leq j \leq N}$ with probability $\{\omega^{(j,t-1)}\}_{1 \leq j \leq N}$.
      Perturb θ to obtain θ^*.
      Simulate from the model 10^3 times according to Algorithm 1.
      Calculate summary statistics, x, from the simulation outputs.
      if $d(D, x) < \epsilon$ **then**
        Accept particle

$$\omega^{(i,t)} \leftarrow \frac{\pi(\theta^{(i,t)})}{\sum_{j=1}^{n} \omega^{(j,t)} K_t(\theta^{i,t} | \theta^{(j,t-1)})}$$

        $i \leftarrow i + 1$
      end if
    end while
  end if
  Normalise weights.
  $t \leftarrow t + 1$
  Set ϵ such that $Pr(g_i \leq \epsilon_i) = q_t$
end while

$$k_f = 2\frac{k_{\text{on}} k_{\text{off}}}{k_{\text{on}} + k_{\text{off}}}$$

To obtain θ, the vector of parameters used in the ABC-SMC inference procedure (Algorithm 2), rate and noise parameters are sampled from the following uniform priors,

$$\begin{aligned}
k_r &\sim U(0, 1) \\
k_f &\sim U(0, 0.1) \\
k_0 &\sim U(0, 1) \\
k_1 &\sim U(0, 1) \\
k_2 &\sim U(0, 10) \\
d_1 &\sim U(0.01, 0.6) \\
d_2 &\sim U(0.0005, 0.05) \\
\eta_{k_{\text{on}}} &\sim U(0, 0.5) \\
\eta_{k_{\text{off}}} &\sim U(0, 0.5) \\
\eta_{k_1} &\sim U(0, 0.4) \\
\eta_{k_2} &\sim U(0, 0.4) \\
\eta_{d_1} &\sim U(0, 0.4) \\
\eta_{d_2} &\sim U(0, 0.4).
\end{aligned}$$

The parameters for the telegraph process, sampled from the prior as k_r and k_f, are converted to k_{on} and k_{off} before being passed to the simulation algorithm (Algorithm 1) as follows,

$$\begin{aligned}
k_{\text{off}} &= \frac{k_f}{2k_r} \\
k_{\text{on}} &= \frac{k_{\text{off}} k_r}{1 - k_r}.
\end{aligned}$$

Rate parameters k_r and k_0 as well as the noise parameters (η) are unit-less. The remaining parameters have units $1s^{-1}$.

To ensure that M and P are from a distribution close to equilibrium, simulation duration is set depending on the nominal degradation rates for mRNA (d_1) and protein (d_2),

$$t_m = L\left(d_1^{-1} + d_2^{-1}\right)$$

where t_m is the final time point and L is a constant chosen arbitrarily to indicate the desired proximity to the steady state distribution. Here we use $L = 5$.

To confirm that our inference procedure is able to converge to the appropriate region of parameter space in an idealised case, we generate synthetic data by simulating 1000 times from the two-state model. We then calculate summary statistics from these data and carry out the inference procedure in the same manner as for the experimental data. Figures 10 and 11 show the resulting distributions of summary statistics and model parameters respectively.

To provide a comparison of the compute times required to simulate the two-state model using the Gillespie algorithm or our model-specific algorithm we take the final population of parameters obtained for the gene dnaK and run simulations on the same CPU using both methods. The extent of the improvement depends on the model parameters. In this case, the mean improvement is 26

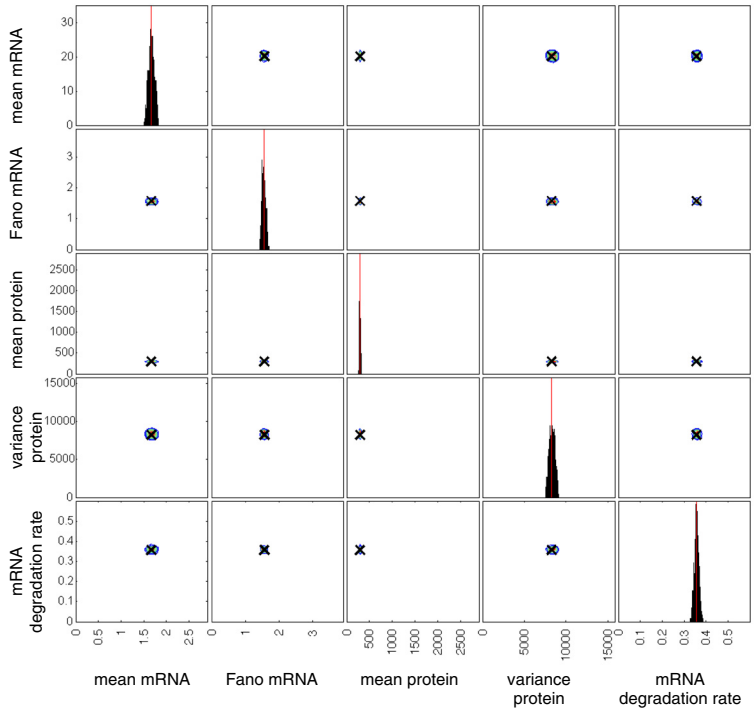

Fig. 10 Posterior distribution of summary statistics and the mRNA degradation rate for a test case where synthetic data were generated by simulating from a model with known parameters. Contour plots indicating the density of points with the corresponding summary statistic for each particle in the final population. The summary statistics for each particle are calculated from 1000 simulation runs. The posterior distribution consists of 1000 particles

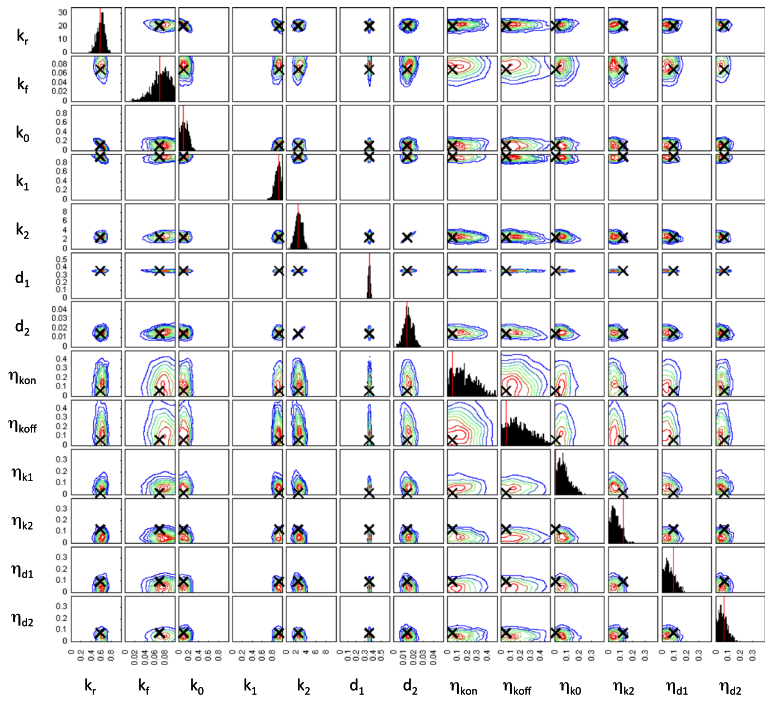

Fig. 11 Posterior distribution of model parameters for a test case where synthetic data were generated by simulating from a model with known parameters. Contour plots indicating the density of points with the corresponding parameter values for each particle in the final population. The posterior distribution consists of 1000 particles

fold with a variance of 12. The total times taken to simulate 1000 perturbed parameter samples from each of 1000 particles were 147 and 3786 s.

Additional files

Additional file 1: Derivation of the Poissonian relationship between the number of surviving protein molecules and mRNA lifetime. (PDF 146 kb)

Additional file 2: Parameter posteriors for the expression model of the *rcsB* gene. (PDF 180 kb)

Additional file 3: Parameter posteriors for the expression model of the *yiiU* gene. (PDF 179 kb)

Additional file 4: Parameter posteriors for the expression model of the *yebC* gene. (PDF 160 kb)

Additional file 5: Parameter posteriors for the expression model of the *eno* gene. (PDF 138 kb)

Abbreviations
ABC, approximate Bayesian computation; SMC, sequential Monte Carlo

Acknowledgements
We thank the members of the *Theoretical Systems Biology Group* at Imperial College London for helpful discussions and feedback.

Funding
The work was supported by a BBSRC Bioprocessing PhD studentship to O.L. and M.P.H.S. and P.D.W.K. was supported by the MRC (project reference MC_UP_ 0801/1).

Availability of data and materials
Previously published data were used [15].

Authors' contributions
OL, PK and MPHS designed the study. OL carried out the computational work. OL and MPHS wrote the paper. All authors read and reviewed the final paper.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]ICR, SM2 5NG Sutton, UK. [2]MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK. [3]Imperial College, London, Centre for Integrative Systems Biology and Bioinformatics, SW7 2AZ London, UK.

Received: 28 October 2015 Accepted: 22 July 2016
Published online: 22 August 2016

References
1. Golding I, Paulsson J, Zawilski SM, Cox EC. Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell. 2005;123(6):1025–36.
2. Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional Pulsing of a Developmental Gene. Curr Biol. 2006;16(10):1018–25.
3. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA Synthesis in Mammalian Cells. PLoS Bio. 2006;4(10):309.
4. Zenklusen D, Larson DR, Singer RH. Single-RNA counting reveals alternative modes of gene expression in yeast. Nat Struct Mol Biol. 2008;15(12):1263–71.
5. Tan RZ, van Oudenaarden A. Transcript counting in single cells reveals dynamics of rDNA transcription. Mol Syst Biol. 2010;6:358.
6. Rosenfeld N. Gene Regulation at the Single-Cell Level. Science. 2005;307(5717):1962–65.
7. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, Emili A, Xie XS. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010;329(5991):533–8.
8. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Sci Adv. 2002;297(5584):1183–186.
9. Spencer SL, Sorger PK, Gaudet S, Albeck JG, Burke JM. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. 2009;459(7245):428–32.
10. Johnston IG, Gaal B, Neves RPd, Enver T, Iborra FJ, Jones NS. Mitochondrial variability as a source of extrinsic cellular noise. PLoS Comput Biol. 2012;8(3):1002416.
11. Kaufmann BB, Yang Q, Mettetal JT, van Oudenaarden A. Heritable stochastic switching revealed by single-cell genealogy. PLoS Biol. 2007;5(9):239.
12. Toni T, Tidor B. Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. PLoS Comput Biol. 2013;9(3):1002960.
13. Zechner C, Ruess J, Krenn P, Pelet S, Peter M, Lygeors J, Koeppl H. Moment-based inference predicts bimodality in transient gene expression. PNAS. 2012;109(21):8340–345.
14. Hasenauer J, Waldherr S, Doszczak M, Radde N, Scheurich P, Allgöwer F. Identification of models of heterogeneous cell populations from population snapshot data. BMC Bioinforma. 2011;12(1):1–15.
15. Cai L, Friedman N, Xie XS. Stochastic protein expression in individual cells at the single molecule level. Nature. 2006;440(7082):358–62.
16. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface. 2008;6(31):187–202.
17. Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MPH. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. Nat Protoc. 2014;9(2):439–56.
18. Karlsson M, Janzen DLT, Durrieu L, Colman-Lerner A, Kjellsson MC, Cedersund G. Nonlinear mixed-effects modelling for single cell estimation: when, why, and how to use it. BMC Syst Biol. 2015;9:52.
19. Raj A, van Oudenaarden A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell. 2008;135(2):216–26.
20. Nienałtowski K, Włodarczyk M, Lipniacki T, Komorowski M. Clustering reveals limits of parameter identifiability in multi-parameter models of biochemical dynamics. BMC Syst Biol. 2015;9(1):65.
21. Erguler K, Stumpf MPH. Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. Mol BioSyst. 2011;7(5):1593.
22. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol. 2007;3(10):1871–878.
23. Secrier M, Toni T, Stumpf MPH. The ABC of reverse engineering biological signalling systems. Mol BioSyst. 2009;5(12):1925.
24. Filippi S, Barnes CP, Cornebise J, Stumpf MPH. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. Stat Appl Genet Mol Biol. 2013;12(1):87–107.
25. Bar-Joseph Z, Gifford DK, Jaakkola TS. Fast optimal leaf ordering for hierarchical clustering. Bioinformatics. 2001;17(Suppl 1):22–9.
26. Kacmar J, Zamamiri A, Carlson R, Abu-Absi NR, Srienc F. Single-cell variability in growing Saccharomyces cerevisiae cell populations measured with automated flow cytometry. J Biotechnol. 2004;109(3):239–54.
27. Yuan TL, Wulf G, Burga L, Cantley LC. Cell-to-Cell Variability in PI3K Protein Level Regulates PI3K-AKT Pathway Activity in Cell Populations. Curr Biol. 2011;21(3):173–83.
28. Li B, You L. Predictive power of cell-to-cell variability - Springer. Quant Biol. 2013;17(1):41–50.
29. Peccoud J, Ycart B. Markovian Modeling of Gene-Product Synthesis. Theor Popul Biol. 1995;48:222–34.
30. Stinchcombe AR, Peskin CS, Tranchina D. Population density approach for discrete mRNA distributions in generalized switching models for stochastic gene expression. Phys Rev E. 2012;85(6):061919.
31. Shahrezaei V, Swain PS. Analytical distributions for stochastic gene expression. Proc Natl Acad Sci. 2008;105(45):17256–17261.
32. Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting systems. J Chem Phys. 2001;115(4):1716–1733.

33. Ale A, Kirk P, Stumpf MPH. A general moment expansion method for stochastic kinetic models. J Chem Phys. 2013;138(17):174101.
34. Lakatos E, Ale A, Kirk P, Stumpf MPH. Multivariate moment closure techniques for stochastic kinetic models. J Chem Phys. 2015;143(9): 094107.
35. Filippi S, Barnes CP, Kirk PDW, Kudo T, Kunida K, McMahon S, Tsuchiya T, Wada T, Kuroda S, Stumpf MPH. Robustness of the MEK-ERK core dynamics and origins of cell-to-cell variability. Cell Rep. 2016;15:2524–535.
36. Lillacci G, Khammash M. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. Bioinformatics. 2013;29(18):2311–319.
37. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys. 1976;22: 403–34.
38. Mc Mahon SS, Lenive O, Filippi S, Stumpf MPH. Information processing by simple molecular motifs and susceptibility to noise. J R Soc Interface. 2015;12(110):20150597.
39. Levsky JM, Singer RH. Fluorescence in situ hybridization: past, present and future. J Cell Sci. 2003;116(Pt 14):2833–838.
40. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
41. Barnes CP, Filippi S, Stumpf M, Thorne T. Considerate approaches to constructing summary statistics for ABC model selection - Springer. Stat Comput. 2012;22(6):1181–197.
42. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interf / R Soc. 2009;6(31):187–202.
43. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. JSTOR: J R Stat Soc Ser B Methodol. 1995;57(1):289–300.