PAPER

cgCorrect: a method to correct for confounding cell–cell variation due to cell growth in single-cell transcriptomics

To cite this article: Thomas Blasi *et al* 2017 *Phys. Biol.* **14** 036001

Physical Biology

PAPER

RECEIVED
2 November 2016

ACCEPTED FOR PUBLICATION
15 February 2017

PUBLISHED
11 May 2017

cgCorrect: a method to correct for confounding cell–cell variation due to cell growth in single-cell transcriptomics

Thomas Blasi[1,2], **Florian Buettner**[1], **Michael K Strasser**[1], **Carsten Marr**[1] **and Fabian J Theis**[1,2]

[1] Institute of Computational Biology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany

[2] Department of Mathematics, Technische Universität München, Garching, Germany

E-mail: fabian.theis@helmholtz-muenchen.de **and** carsten.marr@helmholtz-muenchen.de

Keywords: single cell gene expression, cell growth, computational biology

Supplementary material for this article is available online

Abstract

Accessing gene expression at a single-cell level has unraveled often large heterogeneity among seemingly homogeneous cells, which remains obscured when using traditional population-based approaches. The computational analysis of single-cell transcriptomics data, however, still imposes unresolved challenges with respect to normalization, visualization and modeling the data. One such issue is differences in cell size, which introduce additional variability into the data and for which appropriate normalization techniques are needed. Otherwise, these differences in cell size may obscure genuine heterogeneities among cell populations and lead to overdispersed steady-state distributions of mRNA transcript numbers. We present cgCorrect, a statistical framework to correct for differences in cell size that are due to cell growth in single-cell transcriptomics data. We derive the probability for the cell-growth-corrected mRNA transcript number given the measured, cell size-dependent mRNA transcript number, based on the assumption that the average number of transcripts in a cell increases proportionally to the cell's volume during the cell cycle. cgCorrect can be used for both data normalization and to analyze the steady-state distributions used to infer the gene expression mechanism. We demonstrate its applicability on both simulated data and single-cell quantitative real-time polymerase chain reaction (PCR) data from mouse blood stem and progenitor cells (and to quantitative single-cell RNA-sequencing data obtained from mouse embryonic stem cells). We show that correcting for differences in cell size affects the interpretation of the data obtained by typically performed computational analysis.

1. Introduction

Recent technical advances allow for the analysis of single cells with high throughput omics technologies (Wang *et al* 2010). In particular, single-cell transcriptome analysis (Tang *et al* 2011, Wu *et al* 2014) has made dramatic advances. Investigating transcripts of single cells with both quantitative real-time PCR (qPCR) (Ståhlberg *et al* 2010, Citri *et al* 2012) and single-cell RNA sequencing (RNA-seq) (Tang *et al* 2009, Islam *et al* 2011, 2014, Yan *et al* 2013) has become possible. However, new experimental methods bring new challenges with them: biological variability among single cells, which remained hidden in population-based approaches, has now become evident. One major challenge of computational biology is the development of new and the adaptation of existing methods for single-cell gene expression data (Buettner *et al* 2012, Kim *et al* 2013).

Gene expression is a stochastic process (Elowitz *et al* 2002) and the abundance of mRNA transcripts (of an individual gene) among many single cells (of the same cell type) can be formulated in terms of steady-state probability distributions (Raj *et al* 2006, Thattai *et al* 2001). Analyzing these steady-state probability distributions can yield new insights into the underlying gene expression mechanism (Shahrezaei *et al* 2008, Larson 2011, Kim *et al* 2013).

There are two well-studied mechanisms of gene expression that have been serving as a paradigm (Raj *et al* 2008): simple, constitutive gene expression (also known as the birth-death process), where DNA is continuously transcribed to mRNA (see figure 1(A)); and bursty gene expression, where the DNA promoter successively switches between an active and inactive state and transcripts are produced in episodical bursts (see figure 1(B)). The steady-state distributions of

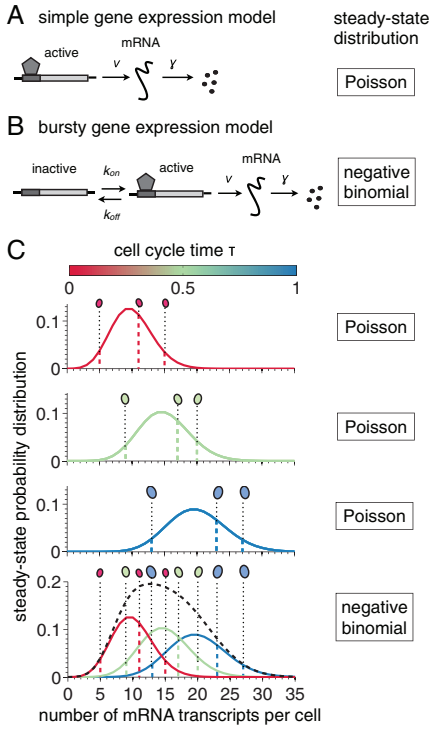

Figure 1. Differences in cell size lead to an overdispersed mRNA distribution. (A) Simple gene expression mechanism. The promoter of a particular gene is always actively transcribing its associated DNA to mRNA with rate ν. The degradation rate of the mRNA is given by γ. (B) Bursty gene expression mechanism. In addition to the simple gene expression mechanism, the promoter can perform transitions between the active and inactive state with rates k_{on} and k_{off}, respectively. (C) A measurement of mRNA transcripts from single cells with different cell sizes that are pooled together leads to an overdispersed steady-state distribution. We display the Poisson distribution of steady-state transcript numbers for three generic cells with different cell sizes (increasing volume from top to bottom) that are all subjected to constitutive, simple gene expression. By pooling these nine cells together and ignoring their different volumes, we obtain an overdispersed steady-state distribution of transcript numbers (bottom panel, dashed black line) that no longer follows a Poisson distribution. The overdispersed distribution may be mistakenly interpreted to be due to the bursty gene expression mechanism.

simple gene expression follow the Poisson distribution (Peccoud *et al* 1995, Thattai *et al* 2001) whereas the steady-state distribution of bursty gene expression follows the negative binomial distribution (Raj *et al* 2006), which allows for more variability among the transcript numbers.

Besides the stochastic nature of gene expression that gives rise to this insightful biological variability, there are also other, confounding sources of variability, such as technical noise (Ramsköld *et al* 2012, Brennecke *et al* 2013, Buettner *et al* 2014, Vallejos *et al* 2015) and cell-cycle effects. The influence of the latter on the interpretation of gene expression data based on steady-state probability distributions has not been investigated so far, even though confounding cell-cycle effects appear in all proliferating cells (such as stem and progenitor cells). During the cell cycle, the cell grows and the number of transcripts within a cell doubles on average (Mitchison 2003). Recently, Padovan-Merhar *et al* (2015) found experimental evidence for the compensation of differences in cell size and suggested that the concentration of transcripts within a cell is maintained constant. This means measuring the abundance of a particular transcript in two identical cells with different cell sizes will yield different results. The differences in cell size cause a broadened, overdispersed steady-state distribution of transcript numbers, which may be mistakenly interpreted in an upstream analysis.

To illustrate this issue, we consider the following scenario (illustrated in figure 1(C)): assume we measure the mRNA transcripts of a particular gene from several single cells, which have the same volume. The gene of interest is subjected to simple, constitutive gene expression and follows the Poisson distribution. In a typical experiment, however, cells are not synchronized and single cells with different sizes are pooled together (see figure 1(C)) leading to an overdispersed steady-state distribution. Performing model selection (see section 2) on the steady-state distribution of transcript numbers obtained by this type of experiment incorrectly favors the negative binomial over the Poisson distribution and therefore the gene expression mechanism would be interpreted to be bursty.

Here, we introduce cgCorrect (cell growth correction), a statistical method to correct single-cell transcriptomics data for latent differences in cell size. cgCorrect can be used for both normalizing single-cell gene expression data sets and for parameter estimation and model selection on steady-state distributions of gene expression. Our approach assumes that the average number of mRNA transcripts within the cell increases proportionally to the volume as the cell grows during the cell cycle, leaving the concentration of transcripts constant (Padovan-Merhar *et al* 2015).

We calculate the cell growth correction probability, which corrects for differences in transcript numbers that are due to differences in cell size. This is the conditional probability of finding the corrected, cell-growth-independent number of mRNA transcripts of a particular gene, given the measured, cell-growth-dependent number of mRNA transcripts of this gene. cgCorrect can include information on the cell volume, but, more strikingly, it can also be applied if there is a total lack of additional information on the cell volume. Since the cell volume is typically not observed, we marginalize this latent variable out, which corresponds to a blind deconvolution problem.

cgCorrect is based on discrete molecule numbers of individual mRNA transcripts in single cells. Discrete molecule numbers are essential for the interpretation of the underlying mechanism of gene expression (Raj *et al* 2009). There are two high throughput transcriptomics techniques, qPCR and RNA-seq, which are both able to measure discrete molecule numbers in single cells (e.g. via digital PCR (Vogelstein *et al* 1999), droplet digital PCR (Hindson *et al* 2011), direct RNA sequencing (Ozsolak *et al* 2009) or strand-specific single-cell sequencing (Islam

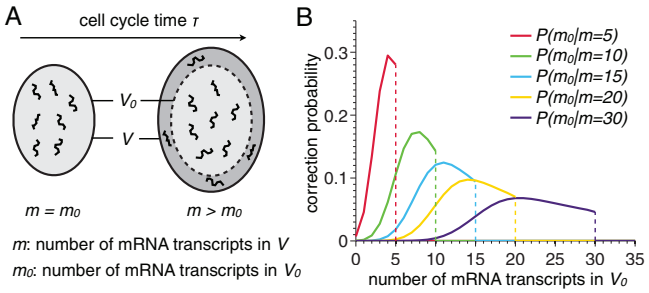

Figure 2. Cell growth model and correction probability. (A) During the cell cycle, a cell will increase its initial volume V_0 to $V > V_0$. We assume the cell to increase its molecular content accordingly, keeping the concentration of mRNA transcripts constant. Then the number of mRNA transcripts m measured later in the cell cycle ($\tau > 0$) is greater than the number of mRNA transcripts m_0 at the beginning of the cell cycle ($\tau = 0$). To render measured mRNA transcript numbers m from single cells at different time points in their cell cycle comparable, we calculate the cell-growth-correction probability $\mathcal{P}_{\mathrm{cgc}}(m_0|m, V)$. This is the probability of finding m_0 transcripts within the cell's initial volume V_0 (light gray area) given the cell's current volume V and the measured number of transcripts m. (B) Cell-growth-correction probability $\mathcal{P}_{\mathrm{cgc}}(m_0|m)$ obtained after marginalizing over a linear growth-model (see text and supplementary figure S1) for several values of measured mRNA transcript numbers m. Notice that $\mathcal{P}_{\mathrm{cgc}}(m_0|m) = 0$ for $m_0 > m$ resulting in the displayed discontinuities.

et al 2011)). If the experiment does not provide discrete molecule numbers, the data can be converted to such by matching the measured value (e.g. cycle time (ct) values in qPCR experiments) to known absolute molecule numbers of a particular gene in the same cell type.

Current state-of-the-art normalization techniques to account for confounding variability are based on scaling the measured number of mRNA transcripts with reporters that should correlate with the confounding variability. In qPCR where the mRNA transcripts of only a few genes are observed, the measured number of transcripts is scaled with the abundance of house-keeping gene transcripts from the same single cell (Guo *et al* 2010, Liviak *et al* 2013, Moignard *et al* 2013). In RNA-seq experiments where the whole transcriptome is measured, the sum of all mRNA transcripts or rank statistics thereof can be used as an estimator for the cell size of each single cell (Brennecke *et al* 2013, Glusman *et al* 2013, Sasagawa *et al* 2013, Vallejos *et al* 2015). However, scaling does not account for the discreteness of mRNA numbers.

Scaling normalization strategies can also be performed based on genes selected from the data, as has been pointed out for bulk measurements (Glusman *et al* 2013). Whereas this approach is infeasible for single cell qPCR, it is applicable for single-cell RNA-seq data since there the whole genome is measured. For instance, it has been shown that the covariance of cell-cycle-related genes can be used to correct for specific gene expression during cell-cycle phases (Buettner *et al* 2015). However, this is not the focus of this work, where we introduce a correction scheme that is based on a global characteristic of each sample, namely the cell volume, rather than on the correlations among the expression of different genes.

2. Methods

2.1. Cell-growth correction probability

Measuring the abundance of a particular mRNA in a single cell during its cell cycle yields a discrete transcript number m, which is generally greater than the transcript number m_0 that we would find at the beginning of the cell's cycle ($\tau = 0$). During the cell cycle, the size of the cell increases from its initial volume $V_0 = V(\tau = 0)$ (at the beginning of its cell cycle) to $V(\tau > 0)$. Cell cycle and cell growth are intimately related (Mir *et al* 2011, Kafri *et al* 2013) and the number of mRNA transcripts within the cell increases as the cell volume increases. Therefore, we assume the concentration of mRNA transcripts m/V to remain constant during the cell cycle. To render the numbers of mRNA transcripts from single cells with different cell sizes comparable, we introduce the volume-dependent cell growth correction probability $\mathcal{P}_{\mathrm{cgc}}(m_0|m, V)$. This is the probability of finding m_0 mRNA transcripts within a cell's initial volume V_0 given a measured number of mRNA transcripts m within a cell's total volume V. The volume-dependent cell growth correction probability is described by a binomial distribution

$$\mathcal{P}_{\mathrm{cgc}}(m_0|m, V) = \mathrm{Bi}(m_0|m, V_0/V), \qquad (1)$$

since this is the discrete probability distribution for finding m_0 transcripts inside the initial volume V_0 given the number of transcripts m present in the total volume V with success rate $p = V_0/V$ (see figure 2(A)). In the limit of high mRNA transcript numbers, the binomial distribution tends to a normal distribution. In this limit, cell growth correction corresponds to scaling the measured number of mRNA transcripts m with the normalized volume of the cell V_0/V. Therefore, the volume-dependent cell growth correction probability, equation (1), contains the commonly performed scaling correction in the limit of high mRNA transcript numbers.

If the single cell's volume V and its initial volume V_0 are measured, we can evaluate $\mathcal{P}_{\mathrm{cgc}}(m_0|m, V)$ directly. In many experimental applications (such as qPCR), however, measuring each single cell's volume is not performed or impossible. In this case, we treat the volume as a latent variable and marginalize over it to obtain the cell growth correction probability

$$\mathcal{P}_{\mathrm{cgc}}(m_0|m) = \int \mathrm{d}V \mathcal{P}_{\mathrm{cgc}}(m_0|m, V)\mathcal{P}(V). \quad (2)$$

To evaluate this, we require the probability distribution of the cells' volumes $\mathcal{P}(V)$ (i.e. the volume distribution over the cell population). This may be determined experimentally or we can use generative models to simulate $\mathcal{P}(V)$ computationally. In the following we used a linear growth model to generate $\mathcal{P}(V)$ (see supplementary material S1 (stacks.iop.org/PhysBio/14/036001/mmedia)). We evaluated the effect of different linear growth models in supplementary figure S2. The cell growth correction probability $\mathcal{P}_{\mathrm{cgc}}(m_0|m)$ for linear growth is displayed in figure 2(B) for several values of observed molecule numbers m.

2.2. cgCorrect for data normalization

The cell growth correction probability $\mathcal{P}_{\mathrm{cgc}}(m_0|m)$ can be used to correct measured mRNA transcript numbers m directly to cell-growth-independent mRNA transcript numbers m_0^* by determining its mode

$$m_0^* = \arg\max_{m_0} \mathcal{P}_{\mathrm{cgc}}(m_0|m). \quad (3)$$

For instance, measuring $m = 15$ transcript numbers in a single cell, the most likely value for the transcript number, which we corrected for differences in cell size is $m_0^* = 11$ (see the blue line in figure 2(B)). This approach offers rank-conserving, one-to-one correspondence between measured and cell-growth-corrected mRNA transcript numbers, as needed for normalization of a data set. When using point estimates (such as the mode of a probability distribution), many alternative mRNA transcript numbers m_0 with non-negligible probability are ignored (see figure 2(B)). However, we can also exploit the full distribution of the correction probability $\mathcal{P}_{\mathrm{cgc}}(m_0|m)$: The number of mRNA transcripts of a particular gene is measured in many single cells. This yields a set of measured mRNA transcript numbers, which we use to obtain the steady-state probability distribution $\mathcal{P}(m)$ of measured mRNA transcript numbers of this gene. We then sum over the correction probability of all measured transcript numbers m multiplied by the steady-state probability distribution to gain the cell-growth-corrected steady-state distribution.

$$\mathcal{P}_{\mathrm{cgc}}(m_0) = \sum_m \mathcal{P}_{\mathrm{cgc}}(m_0|m)\mathcal{P}(m). \quad (4)$$

2.3. cgCorrect for steady-state distribution analysis

The correction probability can also be used to account for differences in cell size when performing steady-state distribution analysis of the transcript numbers of a particular gene. Given the mRNA transcript numbers m of a gene from several single cells, the likelihood $\mathcal{P}(m|\boldsymbol{\theta})$ for the kinetic parameters $\boldsymbol{\theta}$ of the underlying steady-state distribution can be calculated (Peccoud *et al* 1995, Raj *et al* 2006, Shahrezaei *et al* 2008) (see supplementary material S2 for a summary of the analytical steady-state distributions for the simple and bursty gene expression mechanism). Neglecting differences in cell size, however, can lead to incorrect identification of the underlying steady-state distribution and its kinetic parameters (as already demonstrated in figure 1(C)) and has not been considered within this context so far.

Using the correction probability, it is straightforward to incorporate cell growth correction into the existing framework of commonly performed steady-state distribution analysis,

$$\mathcal{P}_{\mathrm{cgc}}(m|\boldsymbol{\theta}) = \sum_{m_0} \mathcal{P}_{\mathrm{cgc}}(m|m_0)\mathcal{P}(m_0|\boldsymbol{\theta}), \quad (5)$$

allowing us to obtain the likelihood for the measured mRNA transcript numbers m from cells that differ in cell size given the parameters $\boldsymbol{\theta}$ of the steady-state distribution under consideration. To obtain $\mathcal{P}_{\mathrm{cgc}}(m|m_0)$ from the correction probability $\mathcal{P}_{\mathrm{cgc}}(m_0|m)$, we use Bayes' theorem with a uniform prior on the measured transcript numbers m.

For the simple gene expression mechanism, the steady-state distribution is given by a Poisson distribution with one kinetic parameter: the effective transcription rate $\lambda = \nu/\gamma$, which corresponds to the mean transcript number among all cells. For bursty gene expression, the steady-state distribution is given by a negative binomial distribution with two kinetic parameters that allow for overdispersion: the burst size $\xi = \nu/k_{\mathrm{off}}$ and the burst frequency $\kappa_{\mathrm{on}} = k_{\mathrm{on}}/\gamma$ (see figure 1 and supplementary material S2 for details on the kinetic parameters). The model parameters can then be found via maximum-likelihood estimation (MLE) $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{P}_{\mathrm{cgc}}(m|\boldsymbol{\theta})$. We evaluate if the parameters of both steady-state distributions are identifiable and therefore capable of describing the data by calculating their profile likelihoods (Raue *et al* 2009).

If both parameters of both distributions are identifiable, we perform model selection using the Bayesian information criterion (BIC) (Jeffreys 1961, Kass *et al* 1995) to select between the Poisson and the negative binomial distribution. Model selection based on the BIC provides a good trade-off between goodness of fit and model complexity; by penalizing models with more parameters it counteracts overfitting the data. In case this model selection is inconclusive ($\Delta\mathrm{BIC} \leqslant 10$), we call the underlying steady-state distribution inconclusive due to model selection (see supplementary material S3 for details on parameter estimation and model selection). As the BIC does not take into account the technical noise level of the data, we only perform model selection for those genes for which the biological variation significantly exceeds technical noise.

2.4. cgCorrect and technical noise correction

In general, cell growth correction can also be combined with technical noise correction. To incorporate technical noise correction into the likelihood, equation (5), the technical noise has to be measured in the experiment (e.g. with external spike-in controls) and the probability distribution of the technical noise

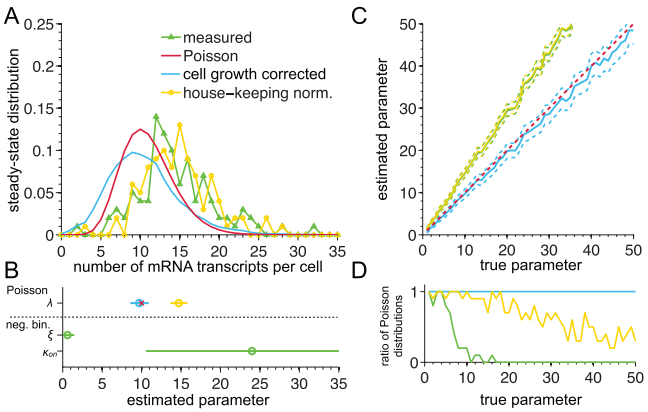

Figure 3. Cell growth correction of simulated gene expression data leads to the correct identification of parameters and the underlying steady-state distribution. (A) Steady-state probability distribution of the measured mRNA transcript numbers m (green line) and the cell growth corrected mRNA transcript numbers m_0 (blue line). The underlying steady-state distribution of transcript numbers in the absence of differences in cell size is given by a Poisson distribution with an effective transcription rate $\lambda_0 = 10$ (red line). The cell growth-corrected probability distribution resembles the Poisson distribution closer than the house-keeping-normalized distribution (yellow line). (B) Estimated parameters and identified steady-state distributions for cell growth-corrected, house-keeping-normalized and measured mRNA transcript numbers. For house-keeping normalization and cell growth correction the steady-state distribution is best described by the Poisson distribution and therefore their gene expression mechanism would be interpreted to be simple; for the measured data, the steady-state distribution is best described by the negative binomial distribution and therefore the gene expression mechanism would be interpreted to be bursty. The Poisson distribution is governed by one parameter: the effective transcription rate $\lambda_0 = \nu/\gamma$. The negative binomial distribution is governed by two parameters: the burst size $\xi = \nu/k_{\text{off}}$ and the burst frequency $\kappa_{\text{on}} = k_{\text{on}}/\gamma$. Only cell growth correction is in the range of the true parameter (red x). Error bars indicate 0.99 confidence intervals of the estimated parameters. (C) To explore the parameter range, we performed parameter estimation and model selection for several values of the effective transcription rate λ_0. We find that cell growth correction (blue line) is capable of correctly inferring the true parameter (red dashed line) for the whole parameter range whereas inferring the parameter on the measured (green line) and the house-keeping-normalized mRNA transcript numbers (yellow line) fails. (D) Ratio of steady-state distributions that were identified to be Poisson for 10 independently simulated data sets. Model selection (based on the BIC) on the cell growth-corrected data (blue line) identifies the true steady-state distribution correctly over the whole parameter range, in contrast to model selection on the measured data (green line) and on the house-keeping-normalized data (yellow line).

$\mathcal{P}_{\text{tn}}(m|m_{\text{t}})$ has to be determined experimentally. This is the conditional probability for the number of mRNA transcripts m that would be measured without technical noise given the number of mRNA transcripts m_{t} that are measured and are subjected to technical noise. The likelihood of cell growth correction and technical noise correction can then be calculated as

$$\mathcal{P}_{\text{cgc,tn}}(m_{\text{t}}|\boldsymbol{\theta}) = \sum_{m} \mathcal{P}_{\text{tn}}(m_{\text{t}}|m)\mathcal{P}_{\text{cgc}}(m|\boldsymbol{\theta}). \quad (6)$$

To obtain $\mathcal{P}_{\text{tn}}(m_{\text{t}}|m)$ from the probability distribution of the technical noise $\mathcal{P}_{\text{tn}}(m|m_{\text{t}})$, Bayes' theorem can be applied with a uniform prior on the measured mRNA transcript numbers with technical noise m_{t}.

If the level of technical noise is higher than the observed level of variability, it is not possible to distinguish between biological variability and variability due to technical noise (Brennecke *et al* 2013, Vallejos *et al* 2015). To assess if the level of observed variability significantly exceeds the level of technical noise we perform an additional test. For each measured gene, we computationally sample transcript measurements from the measured distribution of technical noise for as many cells as were measured in the experiment and calculate their coefficient of variation (CV). We then test for the null hypothesis that the CV of a measured gene is drawn from the sampled distributions of CVs. Only if the null hypothesis is rejected based on a p-value $p \leqslant 0.01$ we use the gene for further analysis with cgCorrect and call it inconclusive due to technical noise otherwise. To correct for multiple testing we used the false discovery rate (Benjamini *et al* 1995).

This test ensures that only genes that have data with a high signal-to-noise ratio enter the subsequent analysis where we estimate parameters and perform model selection on the mRNA steady-state distributions.

3. Results

3.1. cgCorrect on simulated mRNA data leads to correct normalization and identification of the steady-state distribution

To validate cgCorrect, we applied it to mRNA transcript numbers that we simulated from the Poisson distribution (corresponding to the simple gene expression mechanism). We generated mRNA transcript numbers m of 100 single cells with different cell sizes (see supplementary figure S3 and supplementary material S4 for details on the simulation of the data).

Without differences in cell size, the mRNA transcript numbers would be Poisson-distributed $m_0 \sim \text{Pois}(m_0|\lambda_0)$ where the average number of mRNA transcripts per cell $\langle m_0 \rangle = \lambda_0$ equals the effective transcription rate (see red line in figures 1(C) and 3(A) for an example where $\lambda_0 = 10$). Due to differences in cell size, the steady-state distribution of measured mRNA transcript numbers $\mathcal{P}(m)$ is shifted towards higher transcript numbers (green line in figure 3(A)). We

can correct for latent differences in cell size by calculating the corrected steady-state distribution of transcript numbers $\mathcal{P}_{cgc}(m_0)$, equation (4) (blue line in figure 3(A)). Since we ignored the cell volumes by marginalizing the volume out (see equation (2)) the corrected steady-state distribution of transcript numbers does not entirely coincide with the Poisson distribution but has slightly larger tails. To compare cgCorrect with conventional house-keeping normalization, we scaled the measured number of transcripts m with the transcript number of an additionally simulated house-keeping gene m_{hk} (see supplementary material S6 for details on house-keeping normalization), which we chose to have an average number of transcripts $m_{0,hk} = 100$ (yellow line in figure 3(A)). Visual comparison of the two normalization strategies shows that cell growth correction for normalization outperforms house-keeping normalization for this data set.

Model selection based on steady-state distributions reports very strong evidence that the measured steady-state distribution of mRNA transcript numbers can be described by the negative binomial rather than by the Poisson distribution and would therefore mistakenly be interpreted to originate from the bursty gene expression mechanism. When correcting for cell growth, model selection correctly identifies the steady-state distribution to be Poisson (see supplementary material S3 for details on parameter estimation and model selection). By performing parameter estimation, the true effective transcription rate can only be inferred when using cgCorrect (see figure 3(B)), confirming that cgCorrect outperforms house-keeping normalization in recovering the true underlying distribution. To test cgCorrect for a broad parameter range, we simulated additional mRNA data sets for several average numbers of mRNA transcripts per cell (figures 3(C) and (D)): only when we apply cgCorrect are we able to infer the underlying steady-state distribution and its parameters for the whole parameter range correctly.

Moreover, we verified that after applying cgCorrect to transcript numbers that were simulated from the negative binomial distribution, the inferred steady-state distribution is negative binomial. To this end, we simulated mRNA transcript numbers from the negative binomial distribution $m_0 \sim \mathrm{NB}(m_0|\xi, \kappa_{on})$ for a wide range of average numbers of mRNA transcripts $\langle m_0\rangle = \xi \cdot \kappa_{on}$. When applying cgCorrect we found that model selection for $\langle m_0\rangle \geqslant 3$ correctly identifies the underlying steady-state distribution to be negative binomial. For very small average numbers of mRNA transcripts $\langle m_0\rangle \leqslant 2$ the obtained distribution of transcript numbers is very narrow and we find cases (20% for $\langle m_0\rangle = 2$ and 90% for $\langle m_0\rangle = 1$) where the underlying steady-state distribution is identified to be Poisson (see supplementary figure S4). In summary, cgCorrect is capable of both successfully inferring the underlying system parameters from the simulated, cell growth-dependent transcript numbers and correctly specifying the steady-state distribution of transcript numbers.

When analyzing the steady-state distributions of genes, one typically assumes that all cells of a particular cell type share the same kinetic parameters. This assumption does not necessarily reflect biological reality. To explore the effect of neglecting this assumption, we performed simulations where we varied the effective transcription rate of a gene simulated from the simple gene expression mechanism among all cells (see supplementary figure S5). Since the cells' effective transcription rates differ among each other, the simulated steady-state distribution may exhibit overdispersion and model selection may identify the steady-state distributions as being negative binomial.

3.2. cgCorrect on qPCR data facilitates identifying distinct cell types and alters the interpretation of the gene expression mechanism based on a steady-state distribution analysis

To demonstrate the applicability of cgCorrect on single-cell qPCR data, we applied cgCorrect to a recently published data set of hematopoietic stem and progenitor (HSP) cells (Moignard *et al* 2013). In this experiment, 18 transcripts of key hematopoietic genes (and six additional transcripts of house-keeping genes) were measured in 597 single cells of five different HSP cell types. To transform the measured data from ct-values into discrete numbers of mRNA transcript we use results from digital qPCR (Warren *et al* 2006), where the discrete number of one of the 18 transcripts, *PU.1*, was measured for hematopoietic stem cells (HSCs), common lymphoid progenitors (CLPs) and common myeloid progenitors (CMPs), all of them found among the HSPs (see supplementary material S6 for details on the data pre-processing).

Since in this experiment neither technical noise nor information about the cells' volume was measured we applied cgCorrect without technical noise correction and with marginalized volume (equation (2)). To compare cgCorrect with conventional house-keeping normalization, we normalized the data set with the house-keeping genes *Ubc* and *Polr2a* as described by Moignard *et al* (2013). cgCorrect is more suitable to resolve distinct cell types than house-keeping normalization, as can be visualized by PCA (see figure 4). The nearest-neighbor error of finding two differing cell-types next to each other is decreased by 12.1%.

To further illustrate the effect of cgCorrect, we focus on one particular transcript (*PU.1*) in one cell type (CLP) (see figure 5(A)). We analyze the Fano factor $\mathcal{F} = \sigma^2/\mu$ defined as the ratio between the variance σ^2 and the mean μ of the steady-state distribution of mRNA transcript numbers. The Fano factor is a key parameter to quantify deviations from a Poisson distribution (Munsky *et al* 2012) and it equals 1 if the values are Poisson-distributed. cgCorrect alters the Fano factor from $\mathcal{F}(m) = 2.29$ for the measured *PU.1* transcript numbers to $\mathcal{F}(m_0^*) = 1.32$. Parameter estimation for the measured and the corrected transcript numbers is depicted in figure 5(B). cgCorrect alters the identified

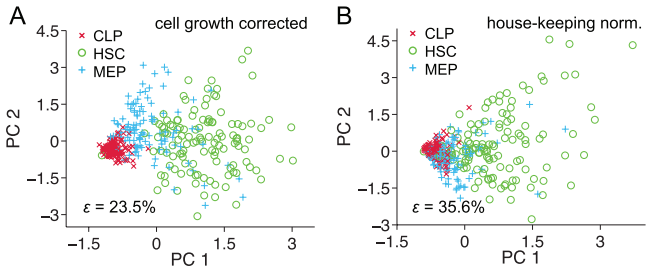

Figure 4. Principal component analysis (PCA) of single-cell qPCR data resolves hematopoietic sub-populations better when using cell growth correction. (A) PCA of cell growth corrected and (B) PCA of house-keeping-normalized single-cell qPCR data of 18 transcripts. The nearest-neighbor error ϵ decreases by 12.1% when using cell growth correction compared to house-keeping normalization.

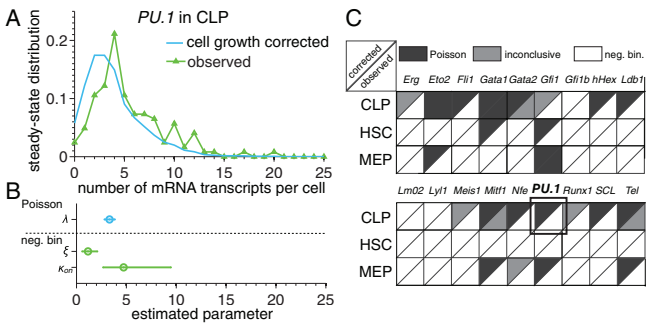

Figure 5. Parameter estimation and model selection on cell growth-corrected steady-state distributions from single-cell qPCR data alters the interpretation of 15 out of the 56 hematopoiesis genes to more likely originate from the simple than from the bursty gene expression mechanism. (A) Probability density of corrected (blue) and measured (green) *PU.1* transcript numbers within CLPs. cgCorrect renders the observed, overdispersed distribution narrower. (B) Estimated kinetic parameters and identified steady-state distribution for *PU.1* mRNA in CLPs. The steady-state distribution is identified to be Poisson after cell growth correction (blue) and negative binomial without cell growth correction (green). cgCorrect alters the interpretation of the underlying gene expression mechanism from bursty to simple. (C) The result of model selection between the Poisson (black) and the negative binomial distribution (white) is visualized. Inconclusive model selection is indicated in gray. The lower right triangle indicates the identified steady-state distribution of the measured transcript numbers, the upper left triangle of the corrected transcript numbers. After performed cgCorrect the identified steady-state distribution is altered in 20 cases.

steady-state distribution of *PU.1* in CLPs from following the overdispersed negative binomial distribution (in case of no correction) to Poisson.

Applying cgCorrect to all measured mRNA transcripts, we find that the steady-state distributions of 18 out of 54 (∼33.0%) gene/cell type combinations are identified to be Poisson and would be interpreted to originate from the simple rather than the bursty gene expression mechanism, whereas this is the case for only three out of 54 (∼5.6%) without cgCorrect (see figure 5(C)). A corresponding analysis with house-keeping normalization yields that the steady-state distribution of only two out of 54 gene/cell type combinations follow the Poisson distribution (see supplementary figure S6).

3.3. cgCorrect on RNA-seq data decreases variability and the number of genes with an overdispersed steady-state distribution

Finally, we applied cgCorrect to a single-cell RNA-seq data set (Islam *et al* 2014) of 41 cells, where 9022 different mRNA transcripts were measured. By using unique molecular identifiers (Kivioja *et al* 2011), it was possible to measure the absolute molecule number of each mRNA within each single cell. Moreover, (Islam *et al* 2014) determined the technical noise with external spike-in control molecules to follow a lossy Poisson distribution—a Poisson distribution with an additional loss factor $f = 0.2$, $\mathcal{P}_{\text{tn}}(m|m_{\text{t}}) = \text{Pois}(m|f \cdot m_{\text{t}})$. Here m_{t} indicates the measured number of mRNA transcripts that are subjected to technical noise and m describes the measured number of mRNA transcripts if there was no technical noise (i.e. the actual number of transcripts in the cell).

We perform steady-state distribution analysis of mRNA transcript numbers of the observed genes and identify the underlying distribution using model selection based on the BIC. In total, 3227 genes significantly ($p \leqslant 0.01$) exceed the limit of technical noise (see section 2) and are further analyzed. In figure 6 the coefficient of variation $\text{CV} = \sigma/\mu$ of the mRNA transcript numbers is displayed as a function of the average mRNA transcript numbers for all measured genes, for the cell growth- and technical noise-corrected (A), for the (only) technical noise-corrected (B) and (C) the uncorrected data set. We indicate the technical noise given by the lossy Poisson distribution as a black solid line. Moreover, we display the limit of Poisson gene expression as a black dashed line.

We find that only a small number of genes, 396, display an overdispersed steady-state distribution

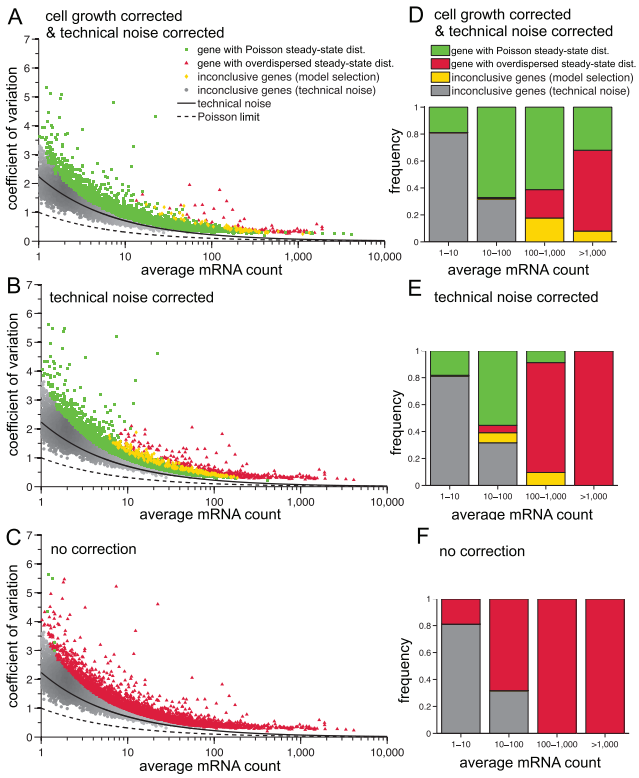

Figure 6. Cell growth correction decreases variability in single-cell RNA-seq data and reduces the number of genes with an overdispersed steady-state distribution from 396 to 85. (A) Scatter plot of the average mRNA transcript numbers versus the coefficient of variation (CV) of all genes in the RNA-seq data set for the cell growth-corrected and technical noise-corrected; (B) only technical noise-corrected and (C) without any correction. Genes with an overdispersed steady-state distribution (red triangles) that would be interpreted to originate from the bursty gene expression mechanism and genes with a Poisson steady-state distribution (green square) that would be interpreted to originate from the simple gene expression mechanism are highlighted. Moreover, we indicate genes with an inconclusive mechanism that is due to model selection (yellow diamond) and due to technical noise (grey circles). The error model for the technical noise, given by the lossy Poisson distribution, is depicted as a solid black line. The limit of Poisson statistics (where the Fano factor $\mathcal{F} = 1$) is indicated with a dashed black line. Overall, we report a shift of the CV towards the Poisson limit when correcting for cell growth. (D) Bar chart of the different identified steady-state distributions from the cell growth and technical noise corrected data; (E) only technical noise-corrected data and (F) uncorrected data. We grouped the genes into logarithmically growing bins. With cell growth correction, in total 396 genes are characterized by an overdispersed steady-state distribution (red area, see supplementary table S2) whereas this is the case for 85 genes (see supplementary table S1) when ignoring differences in cell size and correcting for technical noise only. The majority of the genes' steady-state distributions is Poisson and would be interpreted to originate from the simple gene expression mechanism (green area) or the analysis is inconclusive (yellow and gray area).

when the data is corrected for technical noise. After applying cgCorrect, the number of genes with an overdispersed steady-state distribution is reduced to 85 (see figures 6(A) and (D)). The genes with an overdispersed steady-state distribution after cell growth correction are a subset of the genes with an overdispersed steady-state distribution if only technical noise is corrected (see supplementary tables S1 and S2). Without correction for neither technical noise nor cell-cycle effects, 3222 of these genes display an overdispersed steady-state distribution (see figures 6(C) and (E)).

The steady-state distribution analysis yields an increased proportion of overdispersed genes with a high expression number (>1000 average mRNA count). However, applying cgCorrect to the data decreases this correlation (see supplementary figure S7). When the data is not corrected, almost all genes exhibit overdispersion according to the steady-state distribution analysis. When correcting only for technical noise, all genes with an average mRNA count >1000 are identified to exhibit an overdispersed steady-state distribution. When applying cgCorrect with technical noise correction, the steady-state distribution analysis also identifies highly expressed genes that follow the Poisson distribution.

Since all mRNA transcripts are measured for each single cell, we can also use this additional information for cell growth correction. In this case, we can use the total number of mRNA transcripts within a single cell $M = \sum m$ as an estimator for the cell's volume V. Since the initial volume of the cells V_0 is unknown, we choose the minimal total number of mRNA transcripts among all single cells M_0 as an estimator for the initial volume and therefore set $V_0/V = M_0/M$. We can then apply the volume-dependent correction probability, equation (1), directly without marginalizing over the volume and find that only 32 of the genes display an overdispersed steady-state distribution of transcript numbers (see supplementary figure S8A and C and supplementary table S3). For comparison, we also applied the scaling correction with the total number of mRNA transcripts (Glusman *et al* 2013) (see supplementary

figure S8B and D), which is contained within the limit of large mRNA transcript numbers of the volume-dependent correction probability (see section 2). With the choice of M_0, however, we find a four-fold change among the normalized total number of mRNA transcripts M/M_0 (see supplementary figure S8E) whereas we would expect cell growth to give rise only to a two-fold change. This indicates that using the total number of mRNA transcripts as an estimator for the volume might reduce too much variability.

4. Discussion

In this work, we present cgCorrect, a statistical method for the correction of latent differences in cell size. We show that differences in cell size may lead to an overdispersed steady-state distribution of transcript numbers, which may be misleadingly interpreted in a computational analysis. cgCorrect can be used for data normalization before visualization as well as for steady-state distribution analysis of the data. It can incorporate information about the cell size on different levels: (i) if the size of each cell or an estimator for the size is known, we can use this information to obtain the volume-dependent cell growth correction probability. (ii) If only the probability distribution of the cell volume among the whole population is known, we can use this distribution to marginalize the volume out. (iii) If there is a total lack of information about the cell volume (as is typically the case for qPCR data including the qPCR data set we analyzed), we can use generative growth models to simulate the cell volume distribution computationally and use this for marginalization. Moreover, we showed how cgCorrect can be combined with the correction of technical noise, if the technical noise of the experiment is measured (e.g. via external spike-in controls).

We validated cgCorrect on simulated mRNA data, where we could show that it is only possible to infer the true steady-state distribution and its parameters when cgCorrect was applied. To show that cgCorrect is generally applicable and independent of the experimental setup that was used to measure the data, it was applied on transcriptomics data from qPCR and from RNA-seq. Analyzing steady-state distributions of transcript numbers from a qPCR data set, we found that cgCorrect changed the identified steady-state distribution in 27.4% of the measured cell/gene combinations in HSPs from an overdispersed negative binomial distribution to the Poisson distribution. Moreover, we could show that cgCorrect reduced the number of genes with an overdispersed steady-state distribution in mouse ESCs measured by single-cell RNA-seq from 12.3% to 2.6%. For this data, we could correct for both variability due to technical noise and due to differences in cell size.

In contrast to conventional normalization techniques, cgCorrect takes the discreteness of mRNA transcript numbers into account. For the analyzed qPCR data set, we showed that cgCorrect outperforms traditional house-keeping gene normalization resulting in a better separation of known cell types in PCA. House-keeping genes underlie stochastic gene expression themselves and therefore may not be suitable as reliable reporters for cell size.

In previous analyses, the steady-state distribution of a gene was used to interpret its gene expression mechanism (Raj *et al* 2006, Shahrezaei *et al* 2008, Larson 2011, Kim *et al* 2013). The Poisson steady-state distribution corresponds to the simple gene expression mechanism and the negative binomial distribution corresponds to the bursty gene expression mechanism. However, there are several assumptions involved that are important to consider for this interpretation.

First, it is assumed that the reaction rates that govern the gene expression mechanism remain constant during the cell cycle. Here, we do not consider transcriptional changes during the cell cycle that may alter the reaction rates and have been reported to affect the measured number of mRNA transcripts (Bertoli *et al* 2013, Zopf *et al* 2013). In order to assess the effect of cell cycle-specific gene expression, we modeled transcriptional changes of the reaction rates by an activation function reaching its maximum in the S phase of the cell cycle. The resulting steady-state distribution is identified to follow the overdispersed, negative binomial distribution and would therefore be interpreted to originate from the bursty gene expression mechanism both with and without applying cgCorrect (see supplementary figure S9). Cell cycle-specific gene expression corresponds to a highly orchestrated on and off switching of the promoter region. For a sample of unsynchronized cells that are pooled together, however, the resulting steady-state distribution of mRNA transcript numbers exhibits overdispersion.

The second assumption that is made when analyzing steady-state distribution of mRNA transcript numbers is that the kinetic parameters that govern gene expression are equal for all cells of the same cell type (Thattai *et al* 2001, Raj *et al* 2008, Shahrezaei *et al* 2008, Kim *et al* 2013), which does not necessarily reflect biological reality. We tested the effect on the steady-state distribution analysis when neglecting this assumption by simulating mRNA transcript numbers from a cell population with varying transcription rates expressing mRNAs with the simple mechanism and showed that this effect can also lead to overdispersed steady-state distributions (see supplementary figure S5). A final conclusion on the gene expression mechanism cannot be made based on steady-state distributions of gene expression alone but needs techniques that allow for spatial and temporal resolution such as fluorescence *in situ* hybridization (FiSH) (Raj *et al* 2006, Hocine *et al* 2012, Battich *et al* 2013).

Finally, we made assumptions concerning the cell growth parameters for the generative growth model that we used to obtain the correction probability. The question whether mammalian cells grow linearly or exponentially is still under debate (Cooper 2004,

Popescu *et al* 2014). Here, we used a linear growth model, which has been reported to be appropriate for rat Schwann cells (Conlon *et al* 2003) to computationally simulate the distribution of cell volumes. Moreover, we performed a sensitivity analysis (see supplementary figure S2) that investigates the effect of different linear cell growth scenarios on the correction probability and indicates that our findings are robust with respect to the growth scenario. As already discussed, cgCorrect does not rely on a generative growth model as it allows inclusion of additional information on either each single cell's volume or the distribution of the cells' volume, if they are measured.

To summarize, we identified differences in cell size of proliferating cells to be a latent cause of confounding variability. We introduced cgCorrect, a statistical method that is capable of correcting for this confounding cell-cycle effect in gene expression data, which can be used for data normalization, parameter estimation and model selection. We validated cgCorrect on a simulated data set and applied it to single-cell qPCR gene expression data (Moignard *et al* 2013) from mouse HSPs. Finally, we demonstrated the genome-wide applicability of our approach to single-cell RNA-seq data obtained from mouse ESCs (Islam *et al* 2014).

Acknowledgment

The authors thank John Marioni and Jan Hasenauer for helpful discussions.

This work was supported by the Helmholtz Alliance on Systems Biology (project CoReNe), the European Research Council (starting grant LatentCauses), the Deutsche Forschungsgemeinschaft (SPP 1356 Pluripotency and Cellular Reprogramming) and the Studienstiftung des deutschen Volkes (TB).

References

Battich N *et al* 2013 Image-based transcriptomics in thousands of single human cells at single-molecule resolution *Nat. Methods* **10** 1127–33

Benjamini Y *et al* 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc.* B **57** 289–300

Bertoli C *et al* 2013 Control of cell cycle transcription during G1 and S phases *Nat. Rev. Mol. Cell Biol.* **14** 518–28

Brennecke P *et al* 2013 Accounting for technical noise in single-cell RNA-seq experiments *Nat. Methods* **10** 1093–5

Buettner F *et al* 2012 A novel approach for resolving differences in single-cell expression patterns from zygote to blastocyte *Bioinformatics* **28** i626–32

Buettner F *et al* 2014 Probabilistic PCA of censored data: accounting for uncertainties in the visualisation of high-throughput single-cell qPCR data *Bioinformatics* **30** btu134

Buettner F *et al* 2015 Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells *Nat. Biotechnol.* **33** 155–60

Citri A *et al* 2012 Comprehensive qPCR profiling of gene expression in single neuronal cells *Nat. Protocols* **7** 118–27

Conlon I *et al* 2003 Differences in the way a mammalian cell and yeast cells coordinate cell growth and cell-cycle progression *J. Biol.* **2** 7

Cooper S 2004 Control and maintenance of mammalian cell size *BMC Cell Biol.* **5** 35

Elowitz M *et al* 2002 Stochastic gene expression in a single cell *Science* **297** 1183–6

Glusman G *et al* 2013 Optimal scaling of digital transcriptomes *PLoS One* **8** e77885

Guo G *et al* 2010 Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyte *Dev. Cell* **18** 675–85

Hocine S *et al* 2012 Single-molecule analysis of gene expression using two-color RNA labeling in live yeast *Nat. Methods* **10** 119–21

Hindson B *et al* 2011 High-throughput droplet digital PCR system for absolute quantification of DNA copy number *Anal. Chem.* **83** 8604–10

Islam S *et al* 2011 Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq *Genome Res.* **21** 1160–7

Islam S *et al* 2014 Quantitative single-cell RNA-seq with unique molecular identifiers *Nat. Methods* **11** 163–6

Jeffreys H 1961 *The Theory of Probability* 3rd edn (Oxford: Oxford University Press)

Kafri R *et al* 2013 Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle *Nature* **494** 480–3

Kass R *et al* 1995 Bayes factors *J. Am. Stat. Soc.* **90** 773–95

Kim J K *et al* 2013 Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data *Genome Biol.* **14** R7

Kivioja T *et al* 2011 Counting absolute numbers of molecules using unique molecular identifiers *Nat. Methods* **9** 72–4

Larson D R 2011 What do expression dynamics tell us about the mechanism of transcription? *Curr. Opin. Genet. Dev.* 21 591–9

Liviak K J *et al* 2013 Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells *Methods* **59** 71–9

Mir M *et al* 2011 Optical measurement of cycle-dependent cell growth *Proc. Natl Acad. Sci.* 108 13124–9

Mitichison J M 2003 Growth during the cell cycle *Int. Rev. Cytol.* **226** 165–258 (PMID: 12921238)

Moignard V *et al* 2013 Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis *Nat. Cell Biol.* **15** 363–72

Munsky B *et al* 2012 Using gene expression to understand gene regulation *Science* **336** 183–7

Ozsolak F *et al* 2009 Direct RNA sequencing *Nature* **461** 814–8

Padovan-Merhar O *et al* 2015 Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms *Mol. Cell* **58** 1–14

Popescu G *et al* 2014 New technologies for measuring single cell mass *Lab Chip* **14** 646–52

Peccoud J *et al* 1995 Markovian modeling of gene-product synthesis *Theor. Population Biol.* **48** 222–34

Raj A *et al* 2006 Stochastic mRNA synthesis in mammalian cells *PLoS Biol.* **4** e309

Raj A *et al* 2008 Nature, nurture, or chance: stochastic gene expression and its consequences *Cell* **135** 216–26

Raj A *et al* 2009 Single-molecule approaches to stochastic gene expression *Annu. Rev. Biophys.* **38** 250–70

Ramsköld D *et al* 2012 Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells *Nat. Biotechnol.* **30** 777–82

Raue A *et al* 2009 Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood *Bioinformatics* **25** 1923–9

Sasagawa Y *et al* 2013 Quatz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity *Genome Biol.* **14** R31

Shahrezaei V *et al* 2008 Analytical distributions for stochastic gene expression *Proc. Natl Acad. Sci.* **105** 17256–61

Ståhlberg A *et al* 2010 Single-cell gene expression profiling using reverse transcription quantitative real-time PCR *Methods* **50** 282–8

Tang F *et al* 2009 mRNA-Seq whole-transcriptome analysis of a single cell *Nat. Methods* **6** 377–82
Tang F *et al* 2011 Development and applications of single-cell transcriptome analysis *Nat. Methods* **8** S6–11
Thattai M *et al* 2001 Intrinsic noise in gene regulatory networks *Proc. Natl Acad. Sci.* 98 8614–9
Vallejos C A *et al* 2015 BASiCS: Bayesian analysis of single-cell sequencing data *PLoS Comput. Biol.* **11** e1004333
Vogelstein B *et al* 1999 Digital PCR *Proc. Natl Acad. Sci.* **83** 8604–10
Wang D *et al* 2010 Single cell analysis: the new frontier in 'omics' *Trends Biotechnol.* **28** 281–90
Warren L *et al* 2006 Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR *Proc. Natl Acad. Sci.* 103 17807–12
Wu A *et al* 2014 Quantitative assessment of single-cell RNA-sequencing methods *Nat. Methods* **11** 41–6
Yan L *et al* 2013 Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells *Nat. Struct. Mol. Biol.* **20** 1131–9
Zopf C J *et al* 2013 Cell-cycle dependence of transcription dominates noise in gene expression *PLoS Comput. Biol.* **9** e1003161