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*], Fabian J. Theis[1,2*]

1 Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany.
2 Department of Mathematics, Technische Universität München, Garching, Germany.
∗ **E-mail: fabian.theis@helmholtz-muenchen.de, carsten.marr@helmholtz-muenchen.de**

Abstract

Motivation: Accessing gene expression at the single cell level has unraveled often large heterogeneity among seemingly homogeneous cells, which remained obscured in 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 are differences in cell size, which introduce additional variability into the data, 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.
Results: 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 proportional to the cell's volume during cell cycle. cgCorrect can be used for both data normalization, and to analyze steady-state distributions used to infer the gene expression mechanism. We demonstrate its applicability on both simulated data and single-cell quantitative real-time PCR data from mouse blood stem and progenitor cells. We show that correcting for differences in cell size affects the interpretation of the data obtained by typically performed computational analysis.
Availability: A Matlab implementation of cgCorrect is available at
http://icb.helmholtz-muenchen.de/cgCorrect
Supplementary information: Supplementary information are available online. The simulated data set is available at http://icb.helmholtz-muenchen.de/cgCorrect

1   Introduction

Recent technical advances allow for the analysis of single cells with high throughput omics technologies [Wang *et al.*, 2010]. Especially 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, Yan *et al.*, 2013, Islam *et al.*, 2014] has become possible. But new experimental methods bring new challenges with them: Biological variability among single cells, which remained hidden in population based approaches now becomes evident. One major challenge of computational biology is the development of new and the adaptation of existing methods for single-cell gene expression data [Kim *et al.*, 2013, Buettner *et al.*, 2012].

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 [Thattai *et al.*, 2001, Raj *et al.*, 2006]. Analyzing these steady-state probability distribution 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 birth-death process), where DNA is continuously transcribed to mRNA (see Figure 1A) and bursty gene expression, where the promoter of the DNA successively switches between an active and inactive state and transcripts are produced in episodical bursts (see Figure 1B). The steady-state distributions of simple gene expression follows 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. Especially 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 cell cycle, the cell grows and the number of transcripts within a cell doubles on average [Mitichison, 2003]. Recently [Padovan-Merhar *et al.*, 2015] found experimental evidence for the compensation of differences in cell size and suggest that the concentration of transcripts within a cell is maintained constant. This means that 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 1C): 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 1C) leading to an overdispersed steady-state distribution. Performing model selection (see Methods) 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 is based on the assumption that the average number of mRNA transcripts within the cell increases proportionally to the volume as the cell grows during 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, for 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 cells' volume, but, more strikingly, it can also be applied if there is a total lack of additional information on the cell's 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 both hold the ability 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 *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.
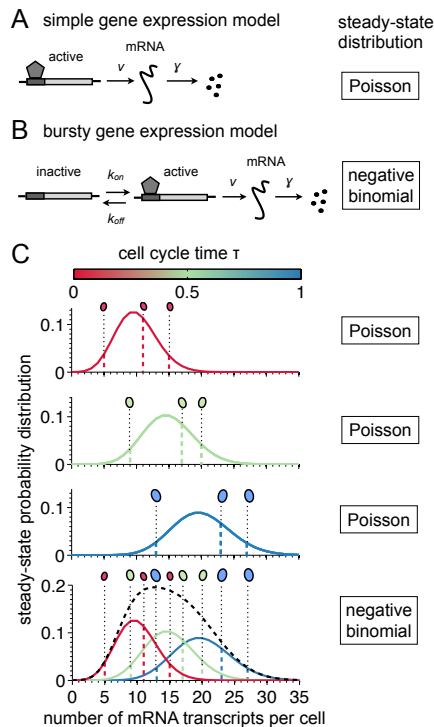
Figure 1. Differences in cell size lead to an overdispersed mRNA distribution. (A) Simple gene expression mechanism. The promoter of a particular gene is always active transcribing its associated DNA to mRNA with rate ν. The degradation rate of the mRNA is given by γ. (B) Bursty gene expression mechanism. Additionally 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 9 cells together and ignoring their different volumes we obtain an overdispersed steady-state distribution of transcript numbers (bottom panel, dashed black line) that does not follow a Poisson distribution any more. The overdispersed distribution may be mistakenly interpreted to be due to the bursty gene expression mechanism.

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 [Moignard *et al.*, 2013, Guo *et al.*, 2010, Liviak *et al.*, 2013]. In RNA-seq experiments where the whole transciptome 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 [Glusman, 2013 *et al.*, Brennecke *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, 2013 *et al.*]. 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 cells' volume, rather than on the correlations among the expression of different genes.

2 Methods

The 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 cell cycle ($\tau = 0$). During 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 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

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

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 2A). 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

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

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
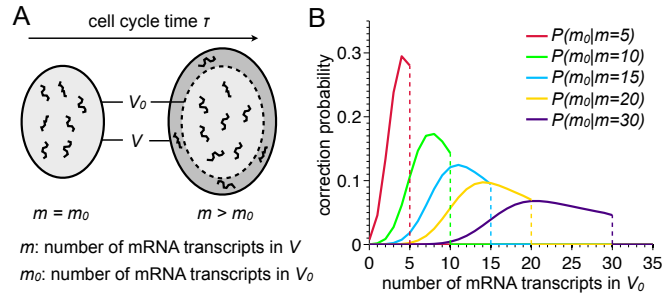
Figure 2. Cell growth model and correction probability. (A) During cell cycle, a cell will increase its initial volume V_0 to $V > V_0$. We assume the cell to increase its molecular content accordingly keeping the concentration of mRNA transcripts constant. Then the number of mRNA transcripts m measured at a latter time in 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 to find m_0 transcripts within the cell's initial volume V_0 (light grey 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.

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). 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 2B for several values of observed molecule numbers m.

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

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

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 blue line in Figure 2B). This approach offers a 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 2B). 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.

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

cgCorrect for steady-state distribution analysis

The correction probability can also be used to account for differences in cell size when performing a 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 1C) 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,

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

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 uniform prior on the measured transcript numbers 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 the 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 Bayesian information criterion (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 ($\Delta\mathrm{BIC} \leq 10$), 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.

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 $\mathcal{P}_{\mathrm{tn}}(m|m_{\mathrm{t}})$ 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

$$\mathcal{P}_{\mathrm{cgc,tn}}(m_{\mathrm{t}}|\boldsymbol{\theta}) = \sum_m \mathcal{P}_{\mathrm{tn}}(m_{\mathrm{t}}|m)\mathcal{P}_{\mathrm{cgc}}(m|\boldsymbol{\theta}). \tag{6}$$

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

3 Results

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 $m_0 \sim$ $\mathrm{Pois}(m_0|\lambda_0)$ where the average number of mRNA transcripts per cell $\langle m_0 \rangle = \lambda_0$ equals the effective transcription rate (see red line in Figure 1C and 3A 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 3A). We can correct for latent differences in cell size by calculating the corrected steady-state distribution of transcript numbers $\mathcal{P}_{\mathrm{cgc}}(m_0)$, Equation (4) (blue line in Figure 3A). Since we ignored the cell's volumes by marginalizing the volume out (cf. 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 $m_{0,\mathrm{hk}} = 100$ (yellow line in Figure 3A). 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 is can be rather described by the negative binomial than by the Poisson distribution and would therefore mistakenly be interpreted to origin 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). Performing parameter estimation the true effective transcription rate can only be inferred when using cgCorrect (see Figure 3B), 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 (Figure 3C and D): Only when we apply cgCorrect we are able to infer the underlying steady-state distribution and its parameters for the whole parameter range correctly.

Moreover, we verified that after applying cgCorrect on 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 $m_0 \sim \mathrm{NB}(m_0|\xi, \kappa_{\mathrm{on}})$ for a wide range of average numbers of mRNA transcripts $\langle m_0 \rangle = \xi \cdot \kappa_{\mathrm{on}}$. When applying cgCorrect we found that model selection for $m_0 \rangle \geq 3$ correctly identifies the underlying steady-state distribution to be negative binomial. For very small average numbers of mRNA transcripts $\langle m_0 \rangle \leq 2$ 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.

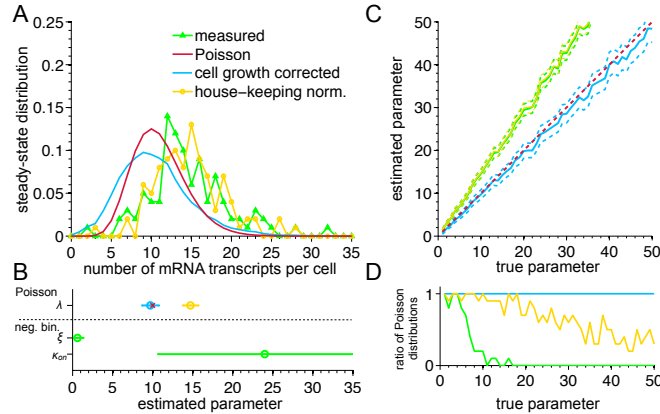When analyzing the steady-state distributions of genes one typically assumes that all cells of a partic-

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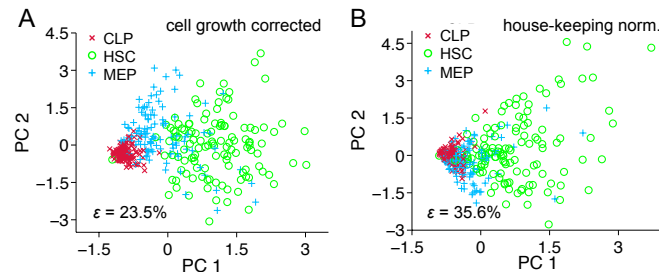
Figure 4. Principal component analysis (PCA) of single-cell qPCR data resolves hematopoietic sub-populations better when using cell growth correction. (A) PCA of cell growth corrected and (B) PCA of house-keeping normalized single-cell qPCR data of 18 transcripts. The nearest neighbor error ϵ decreases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

ular 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.

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 apply 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 better suitable to resolve distinct cell types than house-keeping normalization, as can be visualized by a principal component analysis (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 5A). 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 5B. cgCorrect alters the identified steady-state distribution of *PU.1* in CLPs from following the overdispersed negative binomial distribution (in case of no correction) to Poisson.
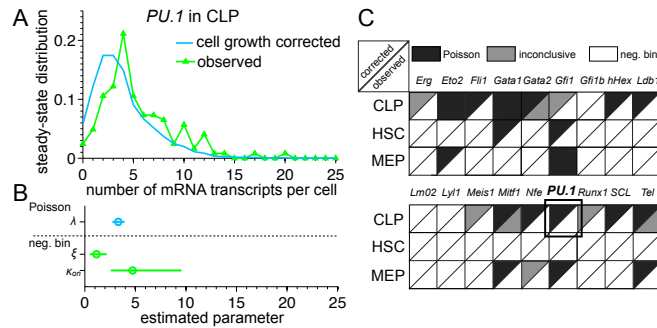
Figure 5. Parameter estimation and model selection on cell growth corrected steady-state distributions from single-cell qPCR data alters the interpretation of 15 out of the 56 hematopoiesis genes to origin more likely from the simple than from the bursty gene expression mechanism. (A) Probability density 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 grey. 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.

Applying cgCorrect to all measured mRNA transcripts, we find that the steady-state distributions of 18 out of 54 ($\sim 33.0\%$) gene/cell type combinations are identified to be Poisson and would be interpreted to origin from the simple rather than the bursty gene expression mechanism, whereas this is the case for only 3 out of 54 ($\sim 5.6\%$) without cgCorrect (see Figure 5C). A corresponding analysis with house-keeping normalization yields that the steady-state distribution of only 2 out of 54 gene/cell type combinations follow the Poisson distribution (see Supplementary Figure S6).

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 a 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 cells' 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 cells' 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 cells' volume distribution computationally and use this for marginalization.

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. 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.

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 a principal component analysis (PCA). House-keeping genes underlie stochastic gene expression themselves and may therefore not suit as reliable reporters for cell size.

In previous analysis the steady-state distribution of a gene is 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 cell cycle. Here, we do not consider transcriptional changes during 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 cell cycle.The resulting steady-state distribution is identified to follow the overdispersed, negative binomial distribution and would therefore be interpreted to origin from the bursty gene expression mechanism both with and without applying cgCorrect (see Supplementary Figure S7). 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, Shahrezaei *et al.*, 2008, Raj *et al.*, 2008, Kim *et al.*, 2013], which does not necessarily have to reflect the 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 to include 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 to correct 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.

Acknowledgement

References

Battich *et al.*, 2013. Image-based transcriptomics in thousands of single human cells at single-molecule resolution, *Nature Methods*, **10**, 1127-1133

Benjamini *et al.*, 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B*, **57**, 289-300.

Bertoli *et al.*, 2013. Control of cell cycle transcription during G1 and S phases, *Nature Reviews Molecular Cell Biology*, **14**, 518-528.

Brennecke *et al.*, 2013. Accounting for technical noise in single-cell RNA-seq experiments, *Nature Methods*, **10**, 1093-1095.

Buettner *et al.*, 2012. A novel approach for resolving differences in single-cell expression patterns from zygote to blastocyte, *Bioinformatics*, **28**, i626-i632.

Buettner *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 *et al.*, 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells, *Nature Biotechnology*, **33**, 155-160.

Citri *et al.*, 2012. Comprehensive qPCR profiling of gene expression in single neuronal cells, *Nature Protocols*, **7**, 118-127.

Conlon *et al.*, 2003. Differences in the way a mammalian cell and yeast cells coordinate cell growth and cell-cycle progression, *Journal of Biology*, **2**, 7.

Cooper, 2004. Cooper,S. (2004) Control and maintenance of mammalian cell size, *BMC Cell Biology*, **5**, 35.

Elowitz *et al.*, 2002. Stochastic Gene Expression in a Single Cell, *Science*, **297**, 1183-1186.

Glusman, 2013 *et al.*. Optimal Scaling of Digital Transcriptomes, *PLOS ONE*, **8**, e77885.

Guo *et al.*, 2010. Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyte, *Developmental Cell*, **18**, 675-685.

Hansen *et al.*, 2012. Allele-specific detection of single mRNA molecules in situ, *Nature Methods*, **10**, 869-871.

Hocine *et al.*, 2012. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast, *Nature Methods*, **10**, 119-121.

Hindson *et al.*, 2011. High-throughput droplet digital PCR system for absolute quantification of DNA copy number, *Analytical Chemistry*, **83**, 8604-8610.

Islam *et al.*, 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, *Genome Research*, **21**, 1160-1167.

Islam *et al.*, 2014. Quantitative single-cell RNA-seq with unique molecular identifiers, *Nature Methods*, **11**, 163-166.

Jeffreys, 1961. The Theory of Probability. Oxford University Press, Oxford, third edition.

Kafri *et al.*, 2013. Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle, *Nature*, **494**, 480-483.

Kass *et al.*, 1995. Bayes factors, *Journal of the American Statistical Society*, **90**, 773-795.

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

Kivioja *et al.*, 2011. Counting absolute numbers of molecules using unique molecular identifiers *Nature Methods*, **9**, 72-74.

Larson, 2011. What do expression dynamics tell us about the mechanism of transcription?, *Current Opinion in Genetics & Development*, **21**, 591-599.

Liviak *et al.*, 2013. Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells, *Methods*, **59**, 71-79.

Mir *et al.*, 2011. Optical measurement of cycle-dependent cell growth, *Proceedings of the National Academy of Sciences*, **108**, 13124-13129.

Mitichison, 2003. Growth During the Cell Cycle, *International Review of Cytology*, **226**, 165-258.

Miyanari *et al.*, 2012. Control of ground-state pluripotency by allelic regulation of Nanog *Nature*, **483**, 470-473.

Moignard *et al.*, 2013. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis, *Nature Cell Biology*, **15**, 363-372.

Munsky *et al.*, 2012. Using Gene Expression to Understand Gene Regulation, *Science*, **336**, 183-187.

Nosrati *et al.*, 2014. Combinatorial action of transcription factors orchestrates cell cycle-dependent expression of ribosomal protein genes and ribosome biogenesis, *FEBS Journal*, **281**, 2339-2352.

Ozsolak *et al.*, 2009. Direct RNA sequencing *Nature*, **461**, 814-818.

Padovan-Merhar *et al.*, 2015. (2015) Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms, *Molecular Cell*, **58**, 1-14.

Popescu *et al.*, 2014. New technologies for measuring single cell mass, *Lab on a Chip*, **14**, 646-652.

Peccoud *et al.*, 1995. Markovian modeling of gene-product synthesis, *Theoretical Population Biology*, **48**, 222-234.

Raj *et al.*, 2006. Stochastic mRNA Synthesis in Mammalian Cells, *PLoS Biology*, **4**, e309.

Raj *et al.*, 2008. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences, *Cell*, **135**, 216-226.

Raj *et al.*, 2009. Single-Molecule Approaches to Stochastic Gene Expression, *Annual Review of Biophysics*, **38**, 250-270.

Ramsköld *et al.*, 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells, *Nature Biotechnology*, **30**, 777-782.

Raue *et al.*, 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, *Bioinformatics*, **25**, 1923-1929.

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

Sass *et al.*, 2014. RAMONA: a Web application for gene set analysis on multilevel omics data, *Bioinformatics*, *Bioinformatics*, bf 31, 128-30.

Shahrezaei *et al.*, 2008. Analytical distributions for stochastic gene expression, *Proceedings of the National Academy of Sciences*, **105**, 17256-17261.

Shapiro *et al.*, 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science, *Nature Reviews Genetics*, **14**, 618-630.

Silvan *et al.*, 2008. Regulation of mRNA Translation during cellular division, *Cell Cycle*, **7**, 741-744.

Ståhlberg *et al.*, 2010. Single-cell gene expression profiling using reverse transcription quantitative real-time PCR, *Methods*, **50**, 282-288.

Tang *et al.*, 2009. mRNA-Seq whole-transcriptome analysis of a single cell, *Nature Methods*, **6**, 377-382.

Tang *et al.*, 2011. Development and applications of single-cell transcriptome analysis, *Nature Methods*, **8**, S6-S11.

Thattai *et al.*, 2001. Intrinsic noise in gene regulatory networks, *Proceedings of the National Academy of Sciences*, **98**, 8614-8619.

Vallejos *et al.*, 2015. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data, *PLoS Computational Biology*, **11**, e1004333.

Vogelstein *et al.*, 1999. Digital PCR *Proceedings of the National Academy of Sciences*, **83**, 8604-8610.

Wang *et al.*, 2010. Single cell analysis: the new frontier in 'omics', *Trends in Biotechnology*, **28**, 281-290.

Warren *et al.*, 2006. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR, *Proceedings of the National Academy of Sciences*, **103**, 17807-17812.

Wu *et al.*, 2014. Quantitative assessment of single-cell RNA-sequencing methods, *Nature Methods*, **11**, 41-46.

Yan *et al.*, 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells, *Nature Structural & Molecular Biology*, **20**, 1131-1139.

Zopf *et al.*, 2013. Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression, *PLoS Computational Biology*, **9**, e1003161.