Discrete Distributional Differential Expression (D^3E) - A Tool for Gene Expression Analysis of Single-cell RNA-seq Data

Mihails Delmans* Martin Hemberg†

July 25, 2015

Abstract

The advent of high throughput RNA-seq at the single-cell level has opened up new opportunities to elucidate the heterogeneity of gene expression. One of the most widespread applications of RNA-seq is to identify genes which are differentially expressed between two experimental conditions. Here, we present a discrete, distributional method for differential gene expression (D^3E), a novel algorithm specifically designed for single-cell RNA-seq data. We use synthetic data to evaluate D^3E, demonstrating that it can detect changes in expression, even when the mean level remains unchanged. Since D^3E is based on an analytically tractable stochastic model, it provides additional biological insights by quantifying biologically meaningful properties, such as the average burst size and frequency. We use D^3E to investigate experimental data, and with the help of the underlying model, we directly test hypotheses about the driving mechanism behind changes in gene expression.

Background

Over the last two decades, several methods for global quantitative profiling of gene expression have been developed [22, 27, 36]. One of the most common uses of gene expression data is to identify differentially-expressed (DE) genes between two samples collected from distinct experimental conditions, e.g.

*Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EA, md656@cam.ac.uk

†Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, mh26@sanger.ac.uk

stimulated vs unstimulated, mutant vs wild-type, or at separate time-points. The goal of DE analysis is to identify genes that underlie the phenotypical differences between the conditions.

The first method for genome-wide expression profiling was microarrays, but as sequencing costs have decreased, direct sequencing of the transcriptome (RNA-seq) has become more popular. Initially, RNA-seq experiments were carried out in bulk on samples of up to 10^5 cells. Consequently, only information about the mean expression of each gene in a sample could be extracted. However, it has been known since the 1950s [21] that gene expression varies from cell to cell, and more recently it has been shown that stochastic variation may play an important role in development, signaling and stress response [25, 26, 38]. Thus, recently developed single-cell RNA-seq protocols [14, 34], could potentially provide a greater understanding of how the transcriptome varies between cells with the same genotype and cell-type. To take full advantage of single-cell data, for DE analysis as well as for other types of investigation, e.g. inference of gene regulatory networks, novel analysis methods are required.

Single-cell DE analysis is complicated by the fact that comparison of two probability distributions is an ambiguous task. With the exception of SCDE [16], most common tools for preforming single-cell DE analysis - DESeq2 [19], Cuffdiff [35], limma [29] and EdgeR [30] - are all adaptations of bulk RNA-sequencing methods. They mainly focus on filtration and normalisation of the raw data, and DE genes are identified based on changes in mean expression levels. The main drawback of using only the mean is that one ignores the gene expression heterogeneity, and will thus fail to detect situations where, for example, there is only a change in the variance of gene expression. Alternative methods for comparing probability distributions are the Kolmogorov-Smirnov test, the Anderson-Darling test, Kullback-Leibler divergence, Akaike's Information Criterion, and the Cramér-von Mises test. What these methods have in common is that they summarize the difference between two distributions as a single value, which can be used to test for significance.

To get the most out of the analysis of single-cell data, one should employ an underlying theoretical model of gene expression. The transcriptional bursting model [23, 24] provides a mechanistic description of the stochastic switching of the promoter as well as the the production and degradation of transcripts at the single cell level (Fig. 1A,B). The model is analytically tractable, and it allows us to derive several other biologically relevant properties of gene expression (Fig. 1C). Despite its simplicity, the transcriptional bursting model enjoys strong experimental support [6, 33, 39, 40].

In this paper, we present D^{3}E, a method based on the comparison of two probability distributions for performing differential gene expression analysis. D^{3}E consists of two separate modules: a module for comparing expression profiles using the Cramér-von Mises, the Anderson-Darling or the Kolmogorov-Smirnov test, and a module for fitting the transcriptional bursting model. Thus, D^{3}E allows the user to go beyond merely identifying DE genes and provides biological insight into the mechanisms underlying the change in expression. We demonstrate the power of D^{3}E to detect changes in gene expression using synthetic data. Finally, we apply D^{3}E to experimental data to demonstrate its ability to detect significant changes which are not reflected by the mean.

Results and Discussion

Algorithm and Implementation

D^{3}E takes a read-count table as an input, with rows and columns corresponding to transcripts and cells, respectively. The user should split the columns into two or more groups by providing cell labels in the input file. If there are more than two groups of cells, they must be compared one pair at a time. D^{3}E uses four steps to process the data. First, input data is normalised using the same algorithm as DESeq2 (see *Methods*) and filtered by removing the genes that are not expressed in any of the cells. Second, the Cramér-von Mises test, the Anderson-Darling test, or the Kolmogorov-Smirnov test [9] is used to identify the genes with a significant change in expression between the two samples of interest. Third, the transcriptional bursting model is fitted to the expression data for each gene in both samples using either the method of moments or a Bayesian method [17]. Fourth, the change in parameters between the two samples is calculated for each gene (Fig. 1D).

A command-line version of D^{3}E written in Python can be downloaded from GitHub (`https://github.com/hemberg-lab/D3E`), and the source code is available under the GPL licence. Furthermore, there is also a web-version available at `http://wwww.sanger.ac.uk/sanger/GeneRegulation_D3E`. Due to the time required to run D^{3}E, the web version limits the number of genes and cells that may be analyzed, and it can only use the method of moments for estimating parameters.

DE Analysis Module

To compare distributions, D^{3}E uses either the Cramér-von Mises, the Anderson-Darling test or the Kolmogorov-Smirnov test to quantify the difference in gene expression (see *Methods*). All three tests are non-parametric which is advantageous since it allows us to apply D^{3}E to any single-cell dataset, not just the ones collected using RNA-seq. The null hypothesis for all three tests is that the two samples are drawn from the same distribution. The premise of D^{3}E is that when two samples are drawn from the same population of cells, the test should result in a high p-value. On the other hand, if the cells are drawn from two populations with different transcriptome profiles, then the resulting p-value should be low. For the remainder of this paper, we will only present results obtained using the Cramér-von Mises test.

We first evaluated D^{3}E using synthetic data. Fortunately, there is a widely used, experimentally validated stochastic model available for single-cell gene expression [23]. We refer to this model as the transcriptional bursting model (Fig. 1A), and it is characterized by three parameters: α, the rate of promoter activation; β, the rate of promoter inactivation; γ, the rate of transcription when the promoter is in the active state. For the purpose of calculating the stationary distribution, all three parameters are normalised by the rate of mRNA degradation λ. Thus, it is only approprite to apply the transcriptional bursting model to DE analysis when λs are constant between the compared samples, or when the degradation rates are known for both samples. The stationary distribution of the transcriptional bursting model takes the form of a Poisson-Beta mixture distribution [24]

$$\begin{aligned} PB(n \mid \alpha, \beta, \gamma) &= \mathrm{Poi}(n \mid \gamma x) \bigwedge_x \mathrm{Beta}(x \mid \alpha, \beta) \\ &= \frac{\gamma^n e^{-\gamma} \Gamma(\alpha + n) \Gamma(\alpha + \beta)}{\Gamma(n+1)\Gamma(\alpha + \beta + n)\Gamma(\alpha)} \Phi(\alpha, \alpha + \beta + n; \gamma), \quad (1) \end{aligned}$$

where n is the number of transcripts of a particular gene, x is an auxiliary variable, Γ is the Euler Gamma function, and $\Phi(a, b; c)$ is the confluent hypergeometric function.

To evaluate the sensitivity of the Cramér-von Mises test to changes in the parameters, we selected triplets of parameters (α, β, γ) from a range that is characteristic for single-cell RNA-seq data. For each parameter triplet one of the parameters was varied, while fixing the remaining two, and a series of Cramér-von Mises tests was carried out on the corresponding Poisson-Beta samples. For each combination of parameters, we assumed that there were 50 cells from each condition when generating the data. The results can be

summarized by a set of matrices, where rows and columns correspond to values of the varied parameter, and the elements in the matrix are p-values from the Cramér-von Mises tests (Fig. 2B). Ideally we would like to find high p-values close to the diagonal and low p-values away from the diagonal. We used a heuristic for characterizing the pattern of p-values, and for each matrix we obtained a single Spearman correlation score, ρ (see *Methods*). A high ρ indicates that the high p-values are concentrated along the diagonal, suggesting that D^{3}E has successfully identified genes where there was a change in one of the parameters.

The results suggest that the Cramér-von Mises test is sensitive to changes in all three parameters with an average score equal to 0.86 (Fig. 2A). The exceptions are the regions in parameter space where γ is small and either β is large or α is small. In this regime, the Poisson-Beta distribution is similar to the Poisson distribution with a mean close to zero, and it is challenging to identify which parameter has changed, and by how much. From a biological perspective, when a transcription rate is small and a gene has a small duty cycle (small α or big β) there are almost no transcripts produced since the promoter spends most of its time in the inactive state. Therefore, changes in either of the three parameters will be difficult to distinguish.

We also considered the scenario when the two distributions are different, but the mean is identical. This is a situation where it is all but impossible for methods which only use the mean to reliably detect that there has been a change in the expression profile. In contrast, D^{3}E is able to reliably identify a change in expression. Our results show that a change of α and β by a factor of 2, which is roughly equivalent to changing the variance by the same factor, is sufficient for the p-value to drop below .05 for a sample of 50 cells (Fig. 2C).

A particular challenge for DE analysis is to determine the p-value threshold for when a change can be considered significant. The traditional approach is to use a fixed value, e.g. .05, and then adjust for multiple hypothesis testing. D^{3}E takes an empirical approach whereby one of the two datasets is first split into two parts. By definition, the two parts should have identical distributions for all genes, which means that it can be used as a negative control. D^{3}E applies the Cramér-von Mises test to the negative control, and records the lowest p-value identified, p^*. When comparing the two original distributions, only genes with a p-value below $a \times p^*$ are considered significant, where the default value for the parameter a is .1.

We used the strategy outlined in the paragraph above, and as a control we generated 1,000 pair of samples with the same number of reads and cells, using identical parameter values for the samples in each pair. We recorded

the lowest observed p-value, p^*, and we used $.1 \times p^*$ as a threshold for when to call a test significant. For both the method of moments and the Bayesian method, we found that 97% of the genes were detected as DE. The control experiment demonstrates that D^{3}E is capable of accurately distinguishing situations where the parameters have truly changed.

To further evaluate the performance of D^{3}E relative to other DE methods, we generated additional synthetic data sets where one of the three parameters was varied while the other two were fixed as before. For each data set we designated genes as DE where the parameter had changed by at least a factor of 1.1, 1.5 or 2. The arbitrary decision of what constitutes a significant change allows us to define the calls of the DE algorithms as either true positive, false positive, true negative or false negative. The results can be summarized as a ROC curve, and again we find that changes in β are more difficult to detect compared to the other two parameters (Fig. 3). We also find that when the threshold for significant changes is set to 10%, then all methods perform at chance level. Importantly, we find that for larger parameter changes, D^{3}E is always amongst the best performing methods (Fig. 3).

Parameter Estimation Module

Even though the rate parameters α, β and γ have well-defined biochemical meanings, they do not represent quantities which can be easily measured. Fortunately, it is possible to use the transcriptional bursting model (Fig. 1) to derive other quantities - the average burst size, the burst frequency, the mean expression level, and the proportion of time in the active state (duty cycle) - which are easier to measure and interpret biologically. How accurately the parameters can be estimated, and how well changes can be identified depends on the sequencing depth. Increasing the sequencing depth will make it easier to identify DE genes, but the total number of reads is typically limited by budgetary constraints. Thus, a key choice when designing an RNA-seq experiment is the trade-off between the total number of cells and the number of reads sequenced per cell.

To determine how the sequencing depth and the number of cells affect our ability to detect DE genes, we generated sets of synthetic data with 1,000 genes in each, while varying the number of cells and the sequencing depth. We evaluated how well D^{3}E is able to identify changes of the inferred parameters in a procedure involving four steps: (i) For each gene, two independent samples were generated by drawing from a Poisson-Beta distributions with separate triplets of randomly generated parameters. The

fold-change of α, β and γ between the two samples was recorded for each gene. (ii) The sequencing procedure was simulated by multinomial sampling (see *Methods*). (iii) The parameters α, β, and γ were estimated by applying either the method of moments or Bayesian inference (see *Methods*) [17] (iv) The performance of the estimation methods was evaluated by comparing the list of genes, sorted based on changes of one of the derived quantities to the ground truth by calculating the normalized number of inversions (see *Methods*). An ideal DE analysis would result in an inversion score equal to 1, i.e. the order in the sorted set of estimated changes for a particular set of parameters would match exactly that of the actual changes. A randomly ordered set of genes would result in an inversion score of 0.5, and a set of genes sorted in reverse order would have an inversion score of zero. The inversion score for both the method of moments and the Bayesian estimates demonstrate a positive trend as the number of cells and the library size increase, although the method of moments performs worse than the Bayesian inference. Comparison of the different quantities reveals that D^3E is best at sorting genes by changes in average burst size and burst frequency (Fig. 4A). Assuming ideal sequencing and 100 cells per sample, the inversion scores are .81 and .82, respectively. The inversion score for duty cycle is always close to 0.5, regardless of sequencing depth or number of cells, suggesting that this quantity is difficult to estimate.

Comparing the benefits of increasing the seqeuncing depth and increasing the sample size, we find that the differences are insignificant when the total number of reads is large. When the total number of reads is small, it is better to have a larger number of cells with shallow sequencing than the other way around. We conclude that the most important recommendation is to avoid operating at the edges of the parameter space. That is, if a small number of cells is used, then the maximum inversion score will be reached sooner than for a larger sample size. For example, if only 10 cells are used, then the maximum inversion score for the burst size is reached with 60 reads per gene. If one instead uses 200 cells, then the maximum inversion for the burst size is reaced with 200 reads per gene (dashed red line in Fig. 4A).

Treatment of zeros

The number of zeros in a read-count table can be large. For example, 72% of entries are zeros in the data reported by [14], 57% for [15], 61 % for [8], 81% for [41], and 52 % for [11]. Zeros originate either from the absence of transcripts in a cell due to stochastic expression, or from technical noise due to low starting levels of mRNA [16]. Thus, removal of zeros may increase

the quality of the analysis by decreasing the technical noise. However, zeros that represent a natural variation would also be removed, which could have an adverse effect on the quality of the results.

To test whether D^{3}E benefits from removing zeros, we generated sets of synthetic data, and randomly substituted a fraction, z, of read-counts with zeros. Next, we applied D^{3}E to the data sets, both before and after having removed the zeros. The effect of removing the zeros was evaluated by the relative change of the inversion score (Fig. 4B). Our results suggest that the effect of removing zeros varies from parameter to parameter, and the results also depend significantly on the magnitude of z. When z is below 25%, estimation of both average burst size and burst frequency suffer significantly from the removal of zeros, as the inversion score drops by $\sim .2$. However, while the effect remains negative for estimation of frequency for $z \geq 25\%$, estimation of average burst size is no longer adversely affected. The estimation accuracy of the duty cycle seems to have improved slightly (up to 10%) after removing zeros. However, taking into account the low predictive power of duty cycle estimates, the absolute value of the improvement is quite modest. In summary, whether or not one should exclude the zeros depends on the estimate of z. According to our simulation studies, it is not advisable to remove zeros when z is below 25%, or when estimating the burst frequency is a high priority.

Application to Experimental Data

The tests on synthetic data suggest that D^{3}E can reliably identify differentially expressed genes. A more useful test of the algorithm, however, involves experimental data which has been reliably validated. Unlike bulk data [27], unfortunately there are no gold-standard datasets available. Nonetheless, to further evaluate D^{3}E, we considered the single-cell RNA-seq data from two and four-cell mouse embryos where qPCR data from the same cell-types was collected for 90 genes [4]. Unfortunately, the correlation of changes in gene expression between the qPCR and RNA-seq data ($\rho_\Delta = .46$) (Fig. S1) is even worse than the correlation of the individual samples ($\rho_2 = .6$, $\rho_4 = .5$). Thus, it does not come as a surprise that the overlap between the genes which are considered DE in the qPCR experiment has little overlap with genes which are considred DE from RNA-seq by any of the five algorithms that we compared (Table S1). Even so, we find large differences in the number of genes identified as DE, ranging from 1 (edgeR) to 35 (DESeq2).

To further evaluate D^{3}E, we applied it to the two datasets collected by Islam *et al.* [14] from 48 mouse embryonic stem cells and 44 mouse

embryonic fibroblasts. To establish the p-value threshold, we first separated the stem cells into two groups, and compared the expression (see *Methods*). We used this approach for determining the threshold for D^{3}E, SCDE, edgeR and limma, while for DESeq2, we used the adjusted p-value reported by the software. When comparing the two cell-types, D^{3}E identified 4197 genes as DE, DESeq2 identified 5183 genes, limma-voom identified 14170 genes, edgeR identified 890 genes, and SCDE identified 1086 genes (Fig. 5A). Surprisingly, the agreement between the five methods is quite low with only a core set of 380 genes identified by all three methods. If we require a gene to be identified of 4 out of 5 methods, then an additional 495 genes are idenitified as DE, suggesting that there is a set of around 900 genes which can confidently considered DE. To further evaluate the set of genes identified as DE by each method, we investigated the distiribution of fold-change values (Fig. 5B). The distributions gives an indication of how large fold changes are required for detection, and we note most of the genes have a higher expression in fibroblasts compared to stem cells. Compared to DESeq2, SCDE and edgeR, we also notice that D^{3}E is able to identify genes with a lower fold change. Indeed, there were several examples of genes where the change in mean expression was modest, but they were still identified by D^{3}E as differentially expressed (Fig. 5C).

Next, we took advantage of the transcriptional bursting model underlying D^{3}E, and we fitted the parameters α, β, and γ for all genes. We found that for 85% of the genes, at least one of the parameters changed by at least 2-fold, suggesting that there are substantial differences between the two cell-types. The results show that all three parameters follow log-normal distributions, spanning approximately one or two orders of magnitude in both cell-types (Fig. 5D). With the exception of the duty cycle which is constrained to be in the interval [0, 1], the derived quantities showed a similar distribution.

Importantly, the transcriptional bursting model allows us to learn more about *how* the expression level has changed between the two conditions. In the transcriptional bursting model, there are three different ways to increase the mean expression level; by decreasing the degradation rate, by increasing the burst frequency, or by increasing the burst size. We calculated the three derived quantities for each condition for the 2105 genes where we were able to obtain degradation rates for both cell-types [31, 32]. Next, we compared the changes in degradation rate, burst frequency and burst size to the change in mean expression level (Fig. 5E). The results clearly demonstrate that it is the change in burst size which underlies the change in mean expression levels ($\rho = .91$), suggesting that altering the burst size is the driving mechanism

behind differences in mean expression between conditions.

Another property of interest is the coefficient of variation (CV), defined as the standard deviation divided by the mean, which is used to quantify the gene expression noise. The CV is inversely correlated with the mean, and the transcriptional bursting model reveals that the change in CV is mainly correlated with the change in the duty cycle ($\rho = .47$), while the effect of a change in burst size is considerably smaller ($\rho = .24$, Fig. S2). To further demonstrate the use of the transcriptional bursting module, we also investigated changes in the auto-correlation times of each gene. The auto-correlation provides information about the time-scale of the noise, i.e. how quickly the gene expression level varies. The expected value of the autocorrelation, τ_c, is given by (*Methods*)

$$\tau_c = \frac{\sigma^2}{\frac{\mu}{\lambda} + \frac{1}{\alpha+\beta}}. \tag{2}$$

Comparison of τ_c and the change in the mean for the Islam *et al* data reveals that the two quantities are strongly correlated ($\rho = .87$, Fig. S3). However, when investigating all the quantities on the right hand side of Eq. (2) the comparison shows that it is the change in variance which is most strongly correlated with the change in autocorrelation times ($\rho = .90$, Fig. S3). Taken together, these results demonstrate that it is possible to generate testable hypotheses about how changes in the property of a gene has come about. The results also show that there is a complex relation between the different properties, and as additional datasets become available, it will be interesting to determine if the correlations observed for the Islam *et al* data can be generalized.

Discussion

DE analysis is one of the most common uses of bulk RNA-seq, and we expect that it will become an important application for single-cell RNA-seq as well. Here, we have presented D^{3}E, a tool for analyszing DE for single-cell data. The main difference between D^{3}E and other methods is that D^{3}E compares the full distribution of each gene rather than just the first moment. Therefore, it becomes possible to identify genes where the higher moments have changed, with the mean remaining constant. To the best of our knowledge, D^{3}E is the first method for DE analysis which takes the full distribution into consideration. Using synthetic data, we demonstrate that D^{3}E can reliably detect when only the shape, but not the mean is changed.

One of the main challenges in developing a DE analysis method for single-cell RNA-seq data is that, unlike for bulk data, there are no gold-standards available [27]. Comparison of qPCR and RNA-seq data revealed only a modest correlation between the two methods, implying that the two methods are inconsistent. Thus, one must resort to synthetic data for evaluation. Fortunately, for single-cell gene expression, there is an analytically tractable transcriptional bursting model available which has been experimentally validated. Even with synthetic data, it is not obvious how one should define a change in expression. Consider the situation where one of the parameters changes by a small amount which is just sufficient to be detected given the limits of the technical noise, the read depth and the sample size. Then the question is whether or not the change is sufficient to be biologically meaningful.

Another challenge stems from the difficulty of disentangling the technical and the biological noise. The transcriptional bursting model does not account for the technical noise in single-cell experiments which can be considerable [3, 5, 11, 28]. Our simulations show, however, that it is possible to improve DE analysis by accounting for the technical noise, demonstrating the importance of estimating the transcript drop-out rate.

D^3E implements three different non-parametric methods for comparing probability distributions. The three methods emphasize different aspects of the distributions, and there are other techniques available for comparing probability distributions. An important future research question is to determine what method is the most appropriate for single-cell DE analysis.

We have shown that the transcriptional bursting model makes it possible to extract additional, biologically relevant results from the DE analysis. However, to be able to fully utilize the transcriptional bursting model, the mRNA degradation rates must be known, or assumed to be constant. Determining degradation rates directly remains experimentally challenging, and today they are only available for a handful of cell-types. However, alternative strategies has been proposed, whereby degradation rates are estimated from RNA-seq data using distribution of reads along the length of a gene [12, 37]. The RNA-seq based methods make it possible to estimate degradation rates without further experiments, and they could thus significantly expand the range of samples where the transcriptional bursting model can be applied.

Conclusions

Our work combines three important aspects of genomics - high-throughput sequencing technologies, computational data analysis, and systems biology modelling. In the present study, we have combined single cell RNA-seq, non-parametric comparison of distributions and an analytical model of stochastic gene expression which allows us to extract biologically meaningful quantities, providing insights not just about which genes have changed between two conditions, but also how the change has come about.

Materials and Methods

Cramér-von Mises criterion

To compare two empirical distributions of read counts from different cell samples, the Cramér-von Mises test was used. Given two discrete distributions $F(x)$ and $G(x)$ with sizes N and M respectively, the Cramér-von Mises test statistic is given by:

$$T = \frac{NM}{N+M} \int_{-\infty}^{\infty} [F(x) - G(x)]^2 H(x), \tag{3}$$

where $H(x)$ is an empirical distribution function of a union of two samples

$$H(x) = \frac{N}{N+M} F(x) + \frac{M}{N+M} G(x). \tag{4}$$

Criterion T was estimated through ranks q_i and s_i of the read-counts from a first and a second samples, in the ordered pooled sample [1]:

$$T = \frac{U}{NM(N+M)} - \frac{4NM+1}{6(N+M)}, \tag{5}$$

where

$$U = N \sum_{i=1}^{N} (q_i - i)^2 + M \sum_{j=1}^{M} (s_i - j)^2. \tag{6}$$

The p-value associated with a null-hypothesis that two samples are drawn from the same distribution was calculated as [2]:

$$\begin{aligned} p(T) &= 1 - \frac{1}{\pi\sqrt{T}} \sum_{j=0}^{\infty} (-1)^j \binom{-0.5}{j} (4j+1)^{0.5} \\ &\quad \exp \frac{-(4j+1)^2}{16T} K_{0.25} \frac{(4j+1)^2}{16T}, \end{aligned} \tag{7}$$

where

$$\binom{-0.5}{j} = \frac{(-1)^j \Gamma(j+0.5)}{\Gamma(0.5) j!}, \tag{8}$$

$\Gamma(z)$ is Euler's Gamma function, and $K_\nu(z)$ is a modified Bessel function of the second kind.

The infinite sum in (7) converges fast after the first few terms. In practice, the p-value was calculated using first 100 terms of the sum for values of T less or equal to 12. For values of T greater than, 12 the p-value was set to zero.

Parameter estimation

A fast but inaccurate method for estimating parameters of a Poisson-Beta distribution is a moments matching technique. The parameters can be expressed through the sample exponential moments [23]:

$$\alpha = \frac{2r_1(r_3 - r_2)}{r_1 r_2 - 2r_1 r_3 + r_2 r_3} \tag{9}$$

$$\beta = \frac{2(r_2 - r1)(r_1 - r_3)(r_3 - r_2)}{(r_1 r_2 - 2r_1 r_3 + r_2 r_3)(r_1 - 2r_2 + r_3)} \tag{10}$$

$$\gamma = \frac{-r_1 r_2 + 2r_1 r_3 - r_2 r_3}{r_1 - 2r_2 + r_3}, \tag{11}$$

where r_i is a successive ratio of exponential moments e_i:

$$r_i = \frac{e_i}{e_{i-1}}, e_0 = 1, \tag{12}$$

for an i'th exponential moment: $e_i = \mathrm{E}[X(X-1)...(X-i+1)]$, where X is a sample of read counts.

The parameters of a Poisson-Beta distribution can also be estimated by a Bayesian inference method [17]. The Bayesian method is more accurate, but it requires more computational power. A Gamma distribution was used as a prior for the parameters α, β and γ:

$$\alpha \sim Gamma(k_\alpha, \theta_\alpha) \tag{13}$$

$$\beta \sim Gamma(k_\beta, \theta_\beta) \tag{14}$$

$$\gamma \sim Gamma(k_\gamma, \theta_\gamma), \tag{15}$$

where

$$k_\alpha = k_\beta = k_\gamma = 1 \tag{16}$$

$$\theta_\alpha = \theta_\beta = 100 \tag{17}$$

$$\theta_\gamma = max\{x : x \in X\}, \tag{18}$$

The number of read counts, x, was drawn from a Poisson-Beta distribution:

$$x \sim Pois(x \mid \gamma c) \bigwedge_{c} Beta(c \mid \alpha, \beta) \tag{19}$$

Parameter estimation was performed by a collapsed Gibbs sampler, using Slice sampling [20]. Conditional distributions for parameters during sampling were given by:

$$P(c_i) \sim Beta(c_i|\alpha,\beta)Poisson(x_i|c_i\gamma) \tag{20}$$

$$P(\alpha) \sim Gamma(\alpha|k_\alpha,\theta_\alpha)\prod_{i=1}^{n} Beta(c_i|\alpha,\beta) \tag{21}$$

$$P(\beta) \sim Gamma(\beta|k_\beta,\theta_\beta)\prod_{i=1}^{n} Beta(c_i|\alpha,\beta) \tag{22}$$

$$P(\gamma) \sim Gamma(\gamma|k_\gamma,\theta_\gamma)\prod_{i=1}^{n} Pois(x_i|c_i\gamma) \tag{23}$$

Synthetic data

Synthetic data was produced by sampling from a Poisson-Beta distribution, i.e. first drawing an auxiliary variable c from Beta distribution with parameters α and β: $c \sim Beta(\alpha, \beta)$ and then drawing from a Poisson distribution with parameter $\lambda = c\gamma$: $x \sim Poisson(c\gamma)$.

The effect of imperfect sequencing was simulated using a Monte Carlo method. A uniformly distributed random variable, v, was drawn from the the interval [0, 1]. Then, a read was assigned to a gene i, where

$$i = \min\{j : C(x_j) \leq v, x \in X\}, \tag{24}$$

where C is a empirical cumulative distribution function for a set of reads for all genes in a particular cell. The operation was performed n times, where n is a total number of reads in a library.

Analyis of the Cramér-von Mises sensitivity

To evaluate how well D^{3}E is capable of detecting changes in different regimes of the parameter space, we systematically varied the three parameters of the Poisson-Beta model across the range of values representative of the biological data, $\alpha \in [.4, 3]$, $\beta \in [2, 100]$, and $\gamma \in [2, 3000]$. We fixed a pair of parameters and varied the third in 10 steps over its range, recording the p-value for

the Cramér-von Mises test. For each of the 100 different combinations, it was assumed that the sample consisted of 50 cells from each condition was generated. Close to the diagonal, the changes in the parameters are small, and we expect a high p-value in these positions. To summarize the matrix of p-values, we calculate the Spearman correlation between the row and column indices where $p > .05$. This value is mapped to a color and reported in Fig. 2A.

Goodness of fit

To evaluate the parameter estimation, the following goodness of fit measure was used. First, a random Poisson-Beta sample with the estimated parameters was generated. The size of the synthetic sample was equal to the size of the real sample. Then the Cramér-von Mises test was performed between the real sample, and the synthetic sample. The p-value of the test is used as a metric for goodness of fit.

Normalization

The normalization of the raw read counts was performed by the same method used by DESeq2 [19]. Let x_{ij} represent the raw number of reads for $i = 1, 2..N$ and $j = 1, 2..M$, where N is the number of genes, and M is the total number of cells in the experiment. Then, the size factor s_j is found as

$$s_j = \text{median}_i \frac{x_{ij}}{(\prod_{k=1}^{M} x_{ik})^{1/M}}. \tag{25}$$

The corrected read counts are then calculated as $x^*_{ij} = \frac{x_{ij}}{sj}$. The size factors are calculated based on spike-ins data only if it is available.

Performance analysis

To test how good D^{3}E is at identifying differences in parameters it was assumed that the power of a DE method depends on it's ability to highlight the biggest changes the in the parameters, i.e. how well it sorts the genes in order of decreasing change in a particular parameter. To quantify this property, we used a normalized number of inversions (inversion score). Assume that $X = \{x_1, x_2...x_n\}$ is a set of numbers sorted descending by some

method. Then, the number of inversions s is given by

$$s(X) = \sum_{i=1}^{n} |\{x_j : x_i \geq x_j, j > i\}| \tag{26}$$

If the set X is perfectly sorted, then it's number of inversions s_0 is $s_0(X) = \sum_{i=1}^{n} i - 1$.

Thus, we can define a normalised number of inversions (inversion score) $s^*(X)$ as $s^*(X) = \frac{s(X)}{s_0(X)}$.

Determining p-value threshold

To determine the p-value threshold for D^{3}E, we first take the sample which will be used as the control group (i.e. in the denominator when calculating the fold-change), and split it into two non-overlapping subsets. Next, the Cramér-von Mises test is applied to the split sample, and the lowest p-value observed, p^*, is recorded. When comparing the case and the control sets, $0.1 * p^*$ is used as a threshold, and only genes with a p-value lower than $0.1 * p^*$ are considered significant.

SCDE reports a z-score which we transform to a p-value using the formula $p = 2\Phi(-|z|)$, where $\Phi(x)$ is the cumulative density of the standard normal distribution. When choosing the threshold for SCDE, we used the same strategy as for D^{3}E.

For DESeq2 we used the adjusted p-value reported by the algorithm, and we required it to be $< .1$ to be significant.

Calculating auto-correlation times

The power spectral density, $S(\omega)$, of the mRNAs for the transcriptional bursting model is given by [7]

$$S(\omega) = \frac{2}{\lambda^2 + \omega^2}\left(d\gamma + \frac{d\alpha\gamma^2}{(\alpha+\beta)^2 - \lambda^2}\right) - \frac{2}{(\alpha+\beta)^2 + \omega^2}\frac{d\alpha\gamma^2}{(\alpha+\beta)^2 - \lambda^2} \tag{27}$$

where $d = \alpha/(\alpha+\beta)$. By definition, the auto-correlation, $R(t)$, is given by the inverse Fourier transform of $S(\omega)$,

$$R(t) = e^{-\lambda|t|}\left(\frac{d\gamma}{\lambda} + \frac{(\alpha+\beta)(d\gamma)^2}{\lambda((\alpha+\beta)^2 - \lambda^2)}\right) - e^{-(\alpha+\beta)|t|}\frac{(d\gamma)^2}{(\alpha+\beta)^2 - \lambda^2}. \tag{28}$$

The characteristic time of the auto-correlation is defined as $\tau_c = S(0)/2R(0)$.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

MH conceived the study and supervised the research, MD and MH carried out the research, MD wrote the code, and MD and MH jointly wrote the manuscript.

Acknowledgements

The authors would like to thank Tallulah Andrews, Daniel Gaffney, Vladimir Kiselev, Michael Kosicki, John Marioni, and Gosia Trynka for helpful discussions and comments on the manuscript. MD was funded by the University of Cambridge BBSRC DTP, and MH was funded by the Wellcome Trust.

References

[1] Anderson, Theodore W., On the Distribution of the Two-Sample Cramér-von Mises Criterion, 1962, *The Annals of Mathematical Statistics 33*, 1148-1159

[2] Anderson, Theodore W., Donald A. Darling, 1952, Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes, *The Annals of Mathematical Statistics 23*, 193-212

[3] Bengtsson, Martin, Martin Hemberg, Patrik Rorsman, Anders Ståhlberg, 2008, Quantification of mRNA in single cells and modelling of RT-qPCR induced noise, *BMC Mol Bio 9:63*, doi:10.1186/1471-2199-9-63

[4] Biase, Fernando H., Xiaoyi Cao, Sheng Zhong, 2014, Cell fate indclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing, *Genome Research 24*, 1787-1796

[5] Brennecke, Philip, Simon Anders, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A. Teichmann, John C. Marioni, Marcus G Heisler, 2013, Accounting for technical noise in single-cell RNA-seq experiments, *Nat. Methods 10*, 1093-95.

[6] Chubb, Jonathan R., Tatjana Trcek, Shailesh M. Shenoy, Robert H. Singer, 2006, Transcriptional Pulsing of a Developmental Gene, *Current Biology 16*, 1018–1025

[7] Coulon, Antoine, Olivier Gandrillon, Guillaume Beslon, 2010, On the spontaneous stochastic dynamics of a single gene: complexity of the molecular interplay at the promoter, *BMC Sys. Bio. 4:2*, doi:10.1186/1752-0509-4-2

[8] Qiaolin, Deng, Daniel Ramsköld, Björn Reinius, Rickard Sandberg, 2014, Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells, *Science 343*, 193-6

[9] Gibbons, Jean D., Subhabrata Chakraborti, 2010, Nonparametric Statistical Inference, 2010, Chapman and Hall

[10] Gillespie, Daniel T., 1976, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, *Journal of Computational Physics 22(4)*, 403-434.

[11] Grün, Dominic, Lennart Kester, Alexander van Oudenaarden, 2014, Validation of noise models for single-cell transcriptomics, *Nat. Methods 11*, 637-40

[12] Gray, Jesse M., David A. Harmin, Sarah A. Boswell, Nicole Cloonan, Thomas E. Mullen, Joseph J. Ling, Nimrod Miller, Scott Kuersten, Yong-Chao Ma, Steven A. McCarroll, Sean M. Grimmond, Michael Springer, 2014, SnapShot-Seq: A Method for Extracting Genome-Wide, In Vivo mRNA Dynamics from a Single Total RNA Sample, *PLoS ONE*, 10.1371/journal.pone.0089673

[13] Holla, M. S., Bhattacharya S. K., 1965, On a discrete compound distribution, *Annals of the Institute of Statistical Mathematics 17*, 377-384

[14] Islam, Saiful, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, Sten Linnarsson, 2011, Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, *Genome Research 11*, 1160-1167

[15] Jääger, Kersti, Saiful Islam, Pawel Zajac, Sten Linnarsson, Toomas Neuman, 2012, RNA-Seq Analysis Reveals Different Dynamics of Differentiation of Human Dermis- and Adipose-Derived Stromal Stem Cells, *PLoS One 7*, e38833

[16] Kharchenko, Peter V, Lev Silberstein, David T Scadden, 2014, Bayesian approach to single-cell differential expression analysis, *Nature Methods 11*, 740-742

[17] Kim, Jong Kyoung, John C. Marioni, 2013, Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data, *Genome Biology 14*, R7

[18] Levine, Joe H., Yihan Lin, Michael B. Elowitz, 2013, Functional Roles of Pulsing in Genetic Circuits, *Science 342*, 1193-1200

[19] Love, Michael I., Wolfgang Huber, Simon Anders, 2014, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biology 15:550*, doi:10.1186/s13059-014-0550-8

[20] Neal, Radford M., 2003, Slice sampling, *The Annals of Statistics*, 705-767

[21] Novick, Aaron, Milton Weiner, 1957, Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci. USA 43*, 553–566

[22] Ozsolak, Fatih, Patrice M. Milos, 2011, RNA sequencing: advances, challenges and opportunities, *Nature Reviews Genetics 12*, 87-98

[23] Peccoud, Jean, Bernard Ycart, 1995, Markovian modelling of gene product synthesis, *Theoretical Population Biology 48*, 222-234

[24] Raj, Arjun, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, Sanjay Tyagi, 2006, Stochastic mRNA Synthesis in Mammalian Cells, *PLoS Biology*, 0.1371/journal.pbio.0040309

[25] Raj, Arjun, Alexander van Oudenaarden, 2008, Stochastic gene expression and its consequences, *Cell 135*, 216–226

[26] Raj, Arjun, Scott A. Rifkin, Erik Andersen, Alexander van Oudenaarden, 2010, Variability in gene expression underlies incomplete penetrance, *Nature 463*, 913-18.

[27] Rapaport, Franck, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E. Mason, Nicholas D. Socci, Doron Betel, 2013, Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data, *Genome Biology*, 14:R95

[28] Risso, Davide, John Ngai, Terence P. Speed, Sandrine Dudoit, 2014, Normalization of RNA-seq data using factor analysis of control genes or samples, *Nat. Biotechnology 32*, 896-902

[29] Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon Smyth, 2015, *limma* powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Research 43(7)*, e47

[30] Robinson, Mark D., Davis J. McCarthy, Gordon K. Smyth, 2010, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics 26*, 139-140

[31] Schwanhäusser, Björn, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, Matthias Selbach, 2011, Global quantification of mammalian gene expression control, *Nature 473*, 337-42

[32] Sharova, Lioudmila V., Alexei A. Sharov, Timur Nedorezov, Yulan Piao, Nabeebi Shaik, Minoru S.H. Ko, 2009, Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells, *DNA Res. 16*, 45-58

[33] Stevense, Michelle, Tetsuya Muramoto, Iris Müller, Jonathan R. Chubb, 2010, Digital nature of the immediate-early transcriptional response, *Development 137*, 579-584

[34] Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, M. Azim Surani, 2009, mRNA-Seq whole-transcriptome analysis of a single cell, *Nat. Methods 6*, 377-82

[35] Trapnell, Cole, David G. Hendrickson, Martin Sauvageau, Loyal Goff, John L. Rinn, Lior Pachter, 2013, Differential analysis of gene regulation at transcript resolution with RNA-seq, *Nature Biotechnology 31*, 46-53

[36] Trevino, Victor, Francesco Falciani, Hugo A Barrera-Saldaña, 2007, DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research, *Molecular Medicine 13*, 527-541

[37] Wan, Lin, Xiting Yan, Ting Chen, Fengzhu Sun, 2012, Modeling RNA degradation for RNA-Seq with applications, *Biostatistics 13*, 734–747

[38] Weinberger, Leor S., John C. Burnett, Jared E. Toettcher, Adam P. Arkin, David V. Schaffer, 2005, Stochastic Gene Expression in a Lentiviral Positive-Feedback Loop: HIV-1 Tat Fluctuations Drive Phenotypic Diversity, *Cell 122(2)*, 169-182.

[39] Wills, Quin F., Kenneth J. Livak, Alex J. Tipping, Tariq Enver, Andrew J. Goldson, Darren W. Sexton, Chris Holmes, 2013, Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments, *Nat. Biotech. 31*, 748-52.

[40] Yunger, Sharon, Liat Rosenfeld, Yuval Garini, Yaron Shav-Tal, 2010, Single-allele analysis of transcription kinetics in living mammalian cells, *Nature Methods 7*, 631-633

[41] Zeisel Amit, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, Sten Linnarsson, 2015, Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq, *Science 347*, 1138-1142

Figures

Additional Files

Additional file 1 — Figure S1

Change in expression levels for 90 genes from the 2-cell and 4-cell mouse embryos as quantified using either qPCR or RNA-seq [4].

Additional file 2 — Table S1

Expression levels for the 90 genes from the 2-cell and 4-cell mouse embryos as quantified using either qPCR or RNA-seq [4]. The last six columns indicate the genes that were identified as differentially expressed by different DE algorithms as well as a t-test for the qPCR data.

Additional file 3 — Parameters for the Islam et al. data without degradation rates

Parameters for the 12,135 genes that were expressed in both cell types.

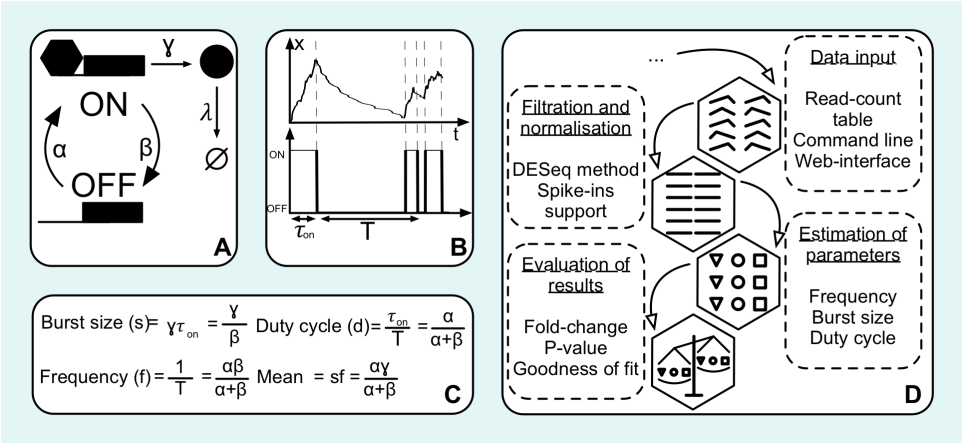

Figure 1: **Overview of D^{3}E. A)** Graphical representation of the transcriptional bursting model. **B)** Example of a realization of the transcriptional bursting model with parameters $\alpha = 1, \beta = 10$, $\gamma = 100$, and $\lambda = 1$ [10]. In this regime, the gene exhibits a bursty behavior with a bimodal stationary distribution. **C)** Derivation of the biologically-relevant parameters from the parameters of the transcriptional bursting model. **D)** Flowchart of the D^{3}E algorithm.

Additional file 4 — Parameters for the Islam et al. data with degradation rates

Parameters for the 2,105 genes that were expressed in both cell types, and where degradation rates were available.

Additional file 5 — Figure S2

Scatterplots showing the mean fold-change, as well as the fold-change of the CV compared to the change in degradation rate, burst frequency, duty cycle, burst size. In all panels, black dots represent genes which did not change, red dots represent genes which were deemed significant by D^{3}E.

Additional file 6 — Figure S3

Scatterplots showing the mean fold-change and the fold-change of the characteristic time, as well as the fold-change of the characteristic time compared to the change in degradation rate, variance and characteristic promoter time.

In all panels, black dots represent genes which did not change, red dots represent genes which were deemed significant by D^3E.

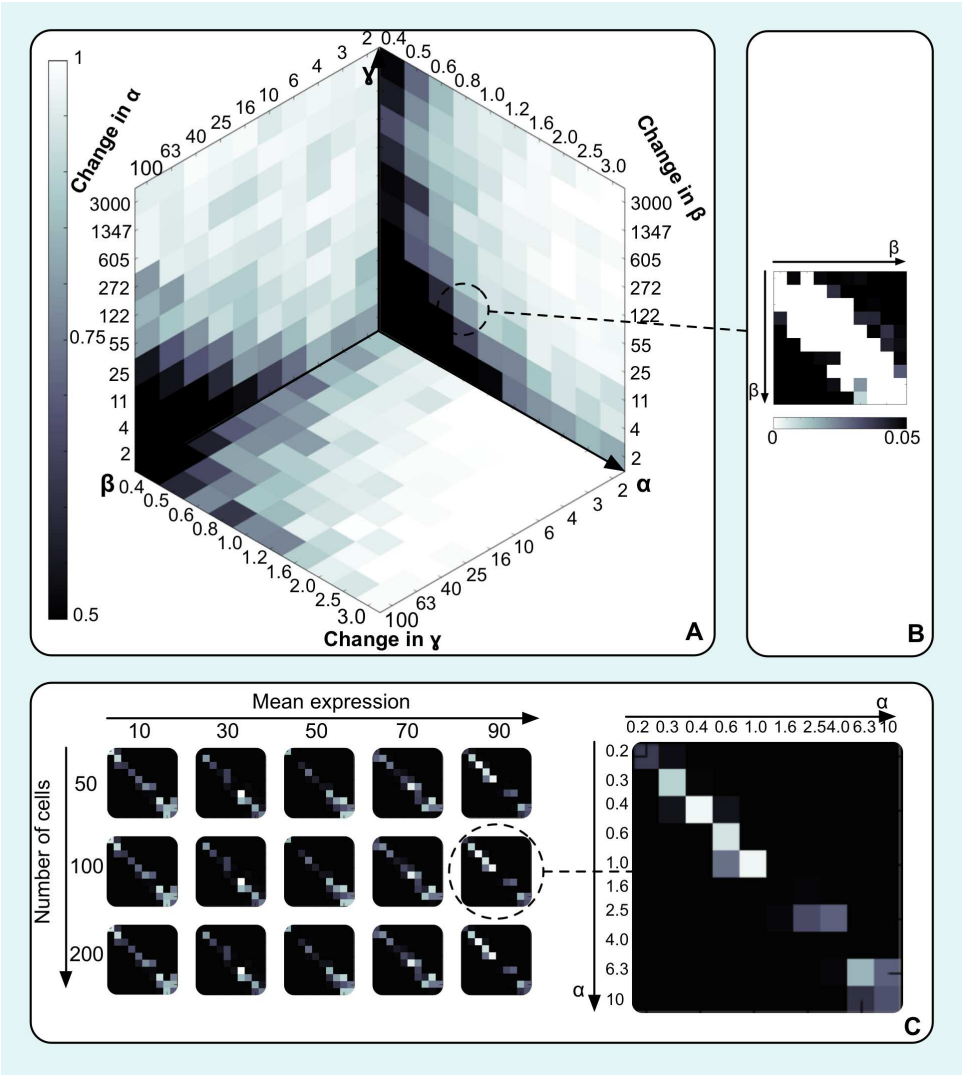

Figure 2: **DE analysis for synthetic data. A)** Cramér-von Mises test sensitivity to changes in parameters of the Poisson-Beta distribution. A lighter color denotes a high sensitivity to changes of a particular parameter. **B)** An example of a matrix which was used to assign the colors in **A**. Here, parameters $\alpha = .8$ and $\gamma = 11$, while β is varied from 0 to 100 on a log-scale. Each element in the matrix reflects a p-value of a Cramér-von Misestest between two Poisson-Beta distributions with the corresponding parameters. We expect to find high values along the diagonal, where the changes are smaller. **C)** DE analysis for the scenario where the mean is fixed but the variance is changed. D^{3}E is able to reliably identify differentially expressed genes based on the change in the shape of distribution alone.

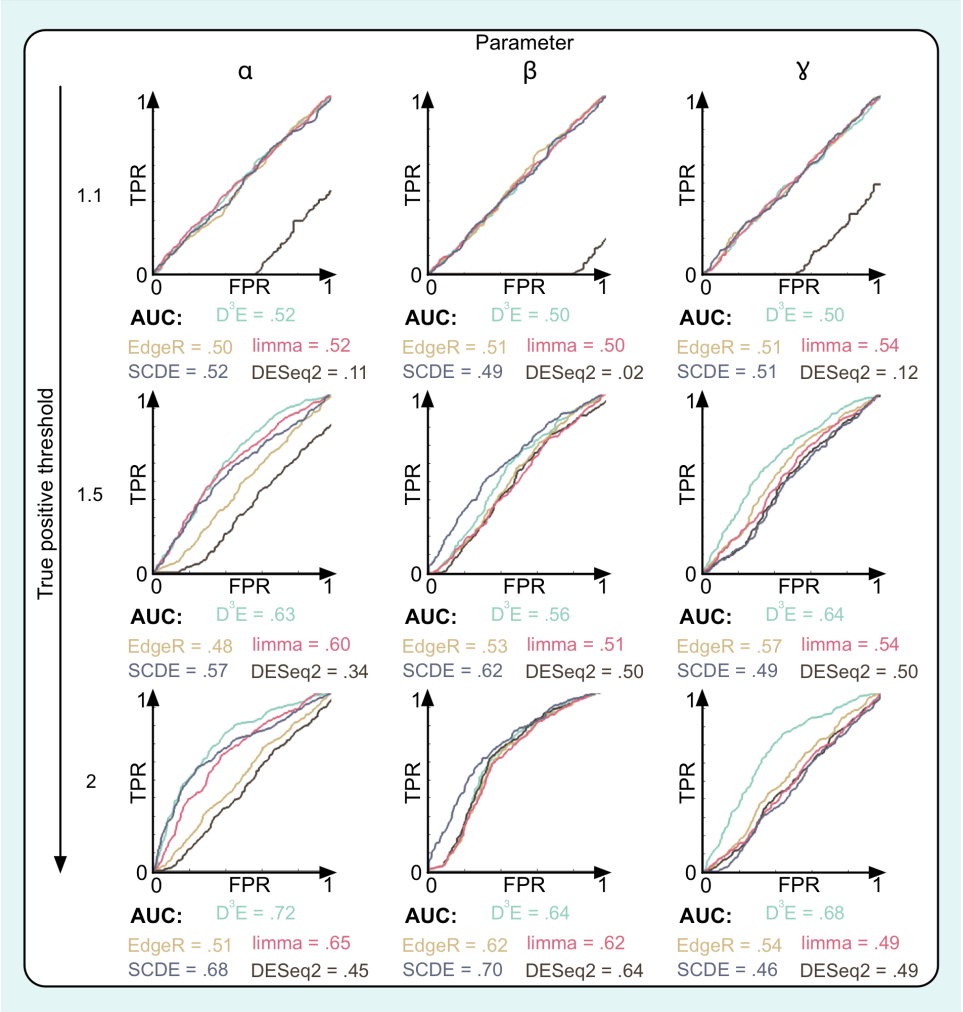

Figure 3: **Comparison of DE methods for synthetic data.** Each panel shows the receiver operator characteristics (ROC) calculated for synthetic data using five different DE algorithms. The numbers below each panel indicate the area under the curve. The rows correspond to different thresholds for when a gene is considered significantly changed. For the first row, DESeq2 reports `NA` for many genes. We treat these calls as false, which explains the unusual shape of the ROC curve.

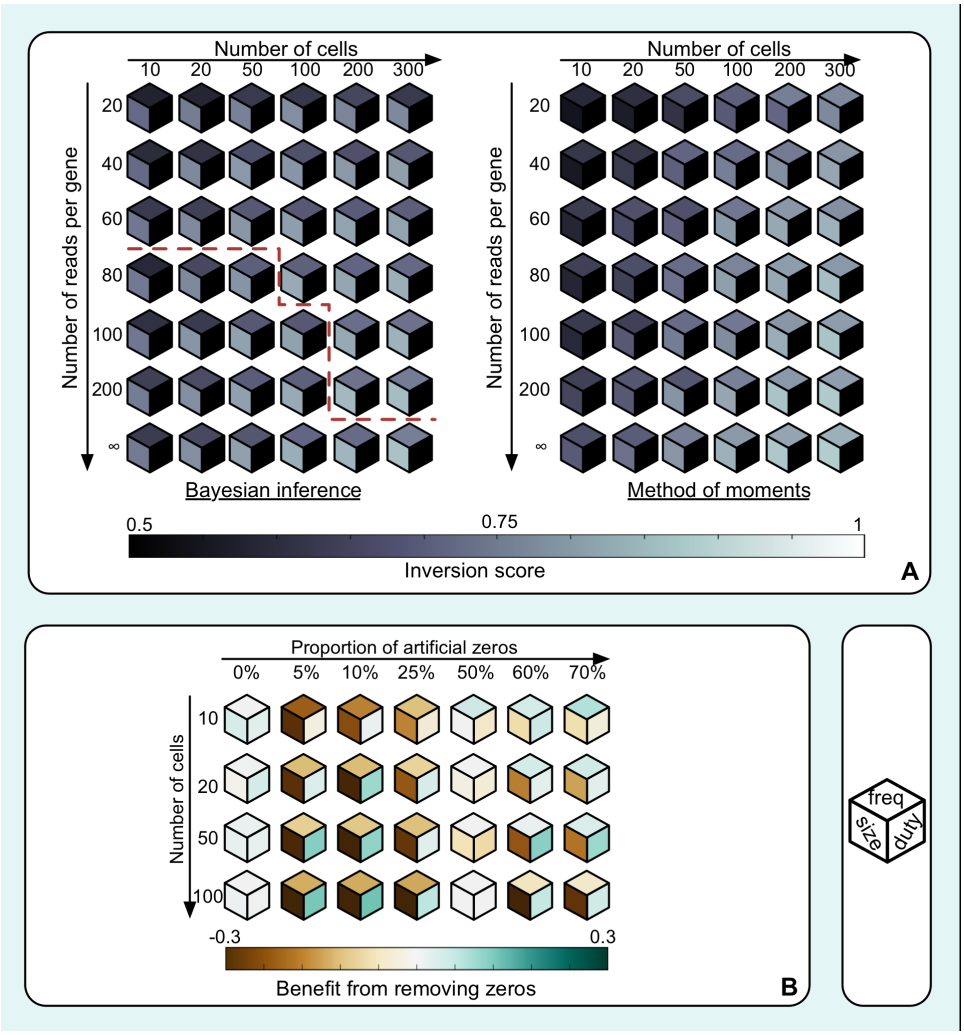

Figure 4: **Detection sensitivity as a function of sample size and sequencing depth. A)** Effect of the number of cells and the depth of sequencing on the performance of D^{3}E analysis using either Bayesian inference or method of moments. The dashed red line represents the sequencing depth where the inversion score for the burst size estimate reaches its maximum for a given number of cells. **B)** The effect of removing zeros from the reads-count table on the performance of D^{3}E analysis is measured by the change in inversion score.

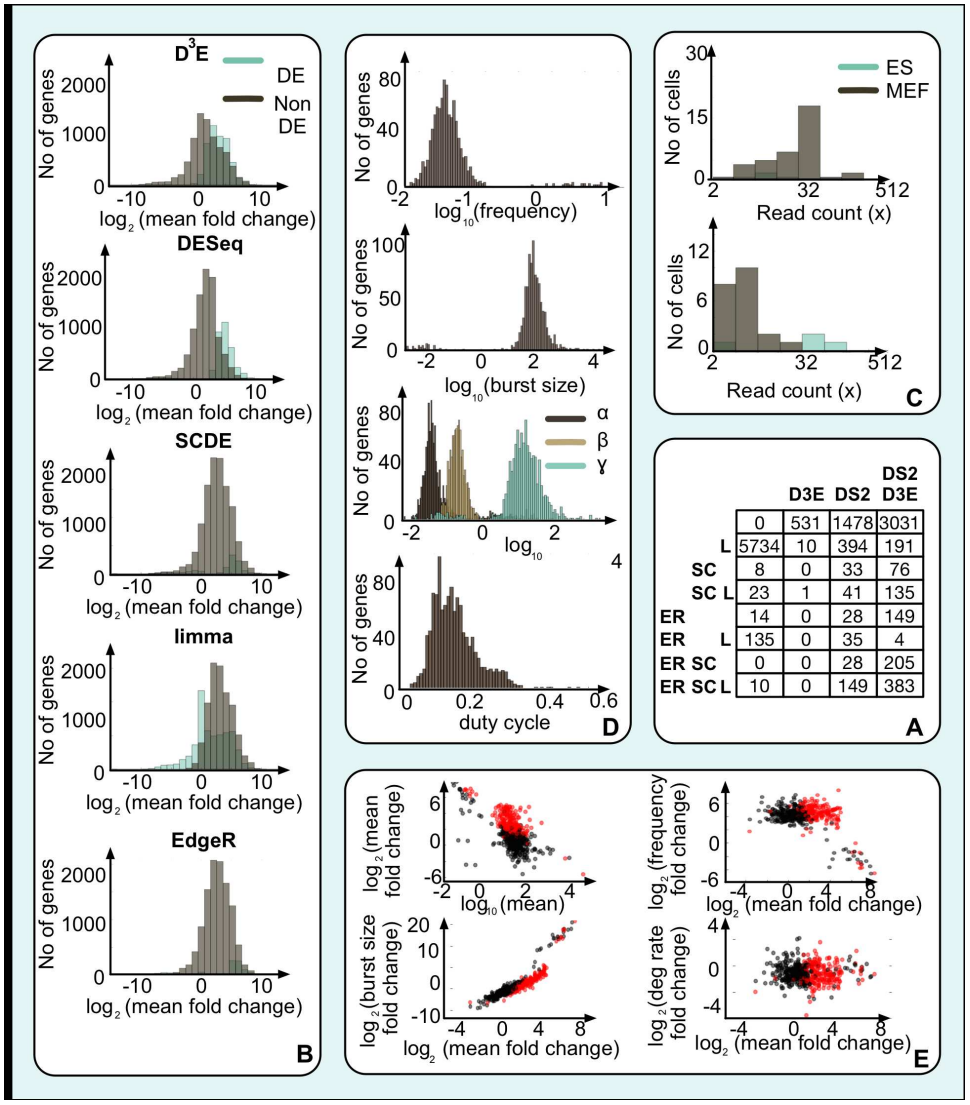

Figure 5: **Analysis of experimental data. A)** Karnaugh table showing the number of genes identified as differentially expressed by D^{3}E, SCDE, limma, edgeR, and DESeq2 for the two datasets collected by Islam *et al* [14]. **B)** Histogram showing the fold-changes for the genes which were considered significantly changed (blue) and not changed (gray) for D^{3}E, DESeq2, limma, edgeR and SCDE. **C)** Examples of two genes, Cdc42bpb in the top panel and Hist1h2bb in the bottom panel, which were identified as DE by D^{3}E. In both cases, the change in mean expression is less than 70% whereas the variance increases by $>$ 10-fold. **D)** Histograms showing the distribution of parameter values for all cells from [14]. From top to bottom, the panels represent the frequency, the burst size, the inferred parameters for the transcriptional bursting model, and the duty cycle. **E)** Scatterplots showing the mean in mESCs, and the fold-change, as well as the fold-change of the mean compared to the change in degradation rate, burst frequency and burst size. In all panels, black dots represent genes which did not change, red dots represent genes which were deemed significant by D^{3}E.