Bayesian approach to single-cell differential expression analysis

Peter V Kharchenko[1–3], Lev Silberstein[3–5] & David T Scadden[3–5]

Single-cell data provide a means to dissect the composition of complex tissues and specialized cellular environments. However, the analysis of such measurements is complicated by high levels of technical noise and intrinsic biological variability. We describe a probabilistic model of expression-magnitude distortions typical of single-cell RNA-sequencing measurements, which enables detection of differential expression signatures and identification of subpopulations of cells in a way that is more tolerant of noise.

Methodological advances are making it possible to examine transcription in individual cells on a large scale[1–4], facilitating unbiased analysis of cellular states[5–8]. However, profiling the low amounts of mRNA within individual cells typically requires amplification by more than 1 million fold, which leads to severe nonlinear distortions of relative transcript abundance and accumulation of nonspecific byproducts. A low starting amount also makes it more likely that a transcript will be 'missed' during the reverse-transcription step and consequently not detected during sequencing. This leads to so-called 'dropout' events, in which a gene is observed at a moderate or high expression level in one cell but is not detected in another cell (**Fig. 1a**). More fundamentally, gene expression is inherently stochastic, and some cell-to-cell variability will be an unavoidable consequence of transcriptional bursts of individual genes or coordinated fluctuations of multigene networks[9]. Such biological variability is of high interest, and several methods have been proposed for detecting it[10–12]. Collectively, this multifactorial variability in single-cell measurements substantially increases the apparent level of noise, posing challenges for differential expression and other downstream analyses.

Comparisons of RNA-seq data from individual cells tend to show higher variability than is typically observed in biological replicates of bulk RNA-seq measurements. In addition to strong overdispersion, there are high-magnitude outliers as well as dropout events (**Fig. 1a**). Such variability is poorly accommodated by standard RNA-seq analysis methods[13,14], and the reported sets of top differentially expressed genes can include high-magnitude outliers or dropout events, showing poor consistency within each cell population (**Fig. 1b**). The abundance of dropout events has been previously noted in single-cell quantitative PCR data and accommodated with zero-inflated distributions[15].

Two prominent characteristics of dropout events make them informative in further analysis of expression state. First, the overall dropout rates are consistently higher in some single-cell samples than in others (**Supplementary Figs. 1** and **2**), indicating that the contribution of an individual sample to the downstream cumulative analysis should be weighted accordingly. Second, the dropout rate for a given cell depends on the average expression magnitude of a gene in a population, with dropouts being more frequent for genes with lower expression magnitude. Quantification of such dependency provides evidence about the true expression magnitude. For instance, dropout of a gene observed at very high expression magnitude in other cells is more likely to be indicative of true expression differences than of stochastic variability.

We modeled the measurement of each cell as a mixture of two probabilistic processes—one in which the transcript is amplified and detected at a level correlating with its abundance and the other in which a transcript fails to amplify or is not detected for other reasons. We modeled the first, 'correlated' component with a negative binomial distribution[13,16]. The RNA-seq signal associated with the second, dropout component could in principle be modeled as a constant zero (i.e., zero-inflated negative binomial process); however, we used a low-magnitude Poisson process to account for some background signal that is typically detected for the dropout and transcriptionally silent genes. Importantly, the mixing ratio between the correlated and dropout processes depends on the magnitude of gene expression in a given cell population. We analyzed two single-cell data sets—a 92-cell set consisting of mouse embryonic fibroblast (MEF) and embryonic stem (ES) cells[2] and a data set of cells from different stages of early mouse embryos[12]. To fit the parameters of an error model for a particular single-cell measurement, we used a subset of genes for which an expected expression magnitude within the cell population can be reliably estimated. Briefly, we analyzed pairs of all other single-cell samples from the same subpopulation (for example, all MEF cells except for the one being fit) with a similarly structured three-component mixture containing one correlated component and dropout components for each cell (**Fig. 1c** and **Supplementary Figs. 1** and **2**). We deemed a subset of genes appearing in correlated components in a sufficiently large fraction of pairwise cell comparisons to be reliable. We estimated the expected expression magnitude of these

[1]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [2]Hematology/Oncology Program, Children's Hospital, Boston, Massachusetts, USA. [3]Harvard Stem Cell Institute, Cambridge, Massachusetts, USA. [4]Center for Regenerative Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA. [5]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to P.V.K. (peter.kharchenko@post.harvard.edu).

RECEIVED 2 OCTOBER 2013; ACCEPTED 28 MARCH 2014; PUBLISHED ONLINE 18 MAY 2014; DOI:10.1038/NMETH.2967

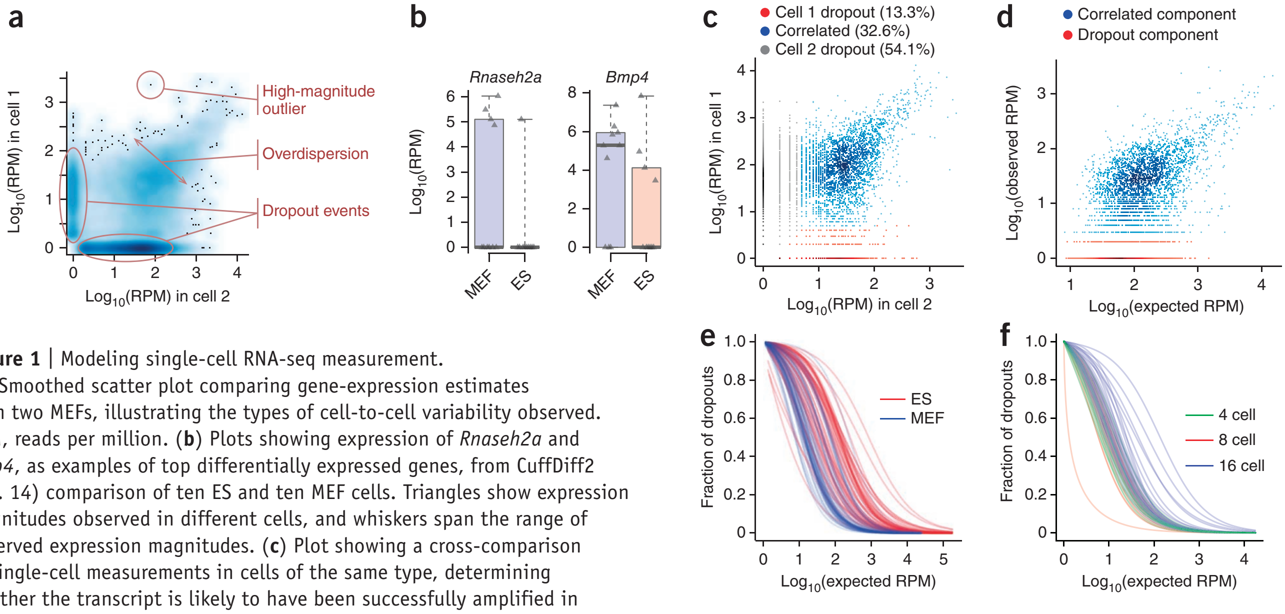

Figure 1 | Modeling single-cell RNA-seq measurement. (**a**) Smoothed scatter plot comparing gene-expression estimates from two MEFs, illustrating the types of cell-to-cell variability observed. RPM, reads per million. (**b**) Plots showing expression of *Rnaseh2a* and *Bmp4*, as examples of top differentially expressed genes, from CuffDiff2 (ref. 14) comparison of ten ES and ten MEF cells. Triangles show expression magnitudes observed in different cells, and whiskers span the range of observed expression magnitudes. (**c**) Plot showing a cross-comparison of single-cell measurements in cells of the same type, determining whether the transcript is likely to have been successfully amplified in both experiments (correlated component). (**d**) Plot showing read counts observed for a particular cell (*y* axis) relative to the expected expression magnitude (*x* axis; see **c**). The measurement is modeled as a mixture of dropout (red) and successful amplification processes (blue), with magnitude-dependent mixing of the two processes. (**e**,**f**) Probability of transcript-detection failures (dropout events) as a function of expression magnitude for individual ES and MEF cells[2] (**e**) and for individual cells from 4-, 8- and 16-cell embryos[12] (**f**).

genes as a median magnitude observed across the cells in which they were found to be part of the correlated components. We then used these expected magnitudes to fit the parameters of the negative binomial distribution as well as the dependency of the dropout rate on the expression magnitude for a given single-cell measurement (**Fig. 1d**). We found that the dropout-rate dependency on the expected expression magnitude can be reliably approximated with logistic regression (**Supplementary Fig. 3**). Notably, the dropout rates vary among cells, depending on the quality of a particular library, cell type or RNA-seq protocol (**Fig. 1e**,**f**).

The error models of individual cells provide a basis for further statistical analysis, for instance to analyze expression differences between groups of single cells. Our Bayesian method for such differential expression analysis (single-cell differential expression, SCDE) incorporates evidence provided by the measurements of individual cells in order to estimate both the likelihood of a gene being expressed at any given average level in each subpopulation and the likelihood of expression fold change between them (**Fig. 2a**,**b**). This approach provides a natural way of integrating uncertain information gained from individual measurements. For example, although an observation of a dropout event in a particular cell does not provide a direct estimate of expression magnitude, it constrains the likelihood that a gene is expressed at high magnitude, in accordance with the overall error characteristics of that cell measurement. To moderate the impact of high-magnitude outlier events, we calculated the joint posterior probability of expression in a cell group by using bootstrap resampling. The resulting sets of top differentially expressed genes can be browsed at http://pklab.med.harvard.edu/scde/. To quantitatively assess the performance of our approach, we evaluated false-positive and false-negative rates based on the expression differences observed in traditional bulk measurements of mouse ES and MEF cells[17] (**Fig. 2c**). The SCDE method shows higher sensitivity than do the common RNA-seq differential expression methods (DESeq and CuffDiff) and the zero-inflated approach developed for quantitative PCR data[15]. The higher SCDE sensitivity was particularly pronounced for genes that are expressed at higher magnitude in ES cells (**Supplementary Fig. 4**), probably owing to the lower total RNA abundance and higher noise levels observed in these cells.

A key promise of the single-cell approach is the ability to discern new subpopulations of cells within complex mixtures in an unbiased manner, without a priori knowledge of which cells are which. Although a variety of existing multivariate analysis techniques can be used to group cells by transcriptional signatures[2,5], dropout and outlier events pose substantial problems for standard similarity measures. Our error models can be used to derive more robust measures. We compared the classification performance of the Pearson linear correlation measure, which has been used in combination with hierarchical clustering to identify transcriptionally distinct subpopulations of cells, with two modified correlation measures that use our error models to account for the likelihood of dropout events. The first measure ('direct dropout') evaluates correlation over a simulated data set in which likely dropout events are designated as missing data. The second ('reciprocal dropout') weights the contribution of each gene on the basis of the probability that the gene will fail (drop out) in the second cell, given its expression level in the first cell (Online Methods). Evaluating the performance of different correlation measures over increasingly difficult cell classification, we found that measures adjusted on the basis of the derived error models perform consistently better in resolving cell populations (**Fig. 2d** and **Supplementary Fig. 5**).

Genome-wide transcriptional examination of cellular heterogeneity within complex tissues will redefine the boundaries separating cellular states in statistical terms[18]. Here we have used a simple mixture model to capture the uncertainty in expression magnitude observed in a given cell, propagating this uncertainty into subsequent analyses. As single-cell studies gain in scope, such probabilistic views of the transcriptional state will become increasingly important.

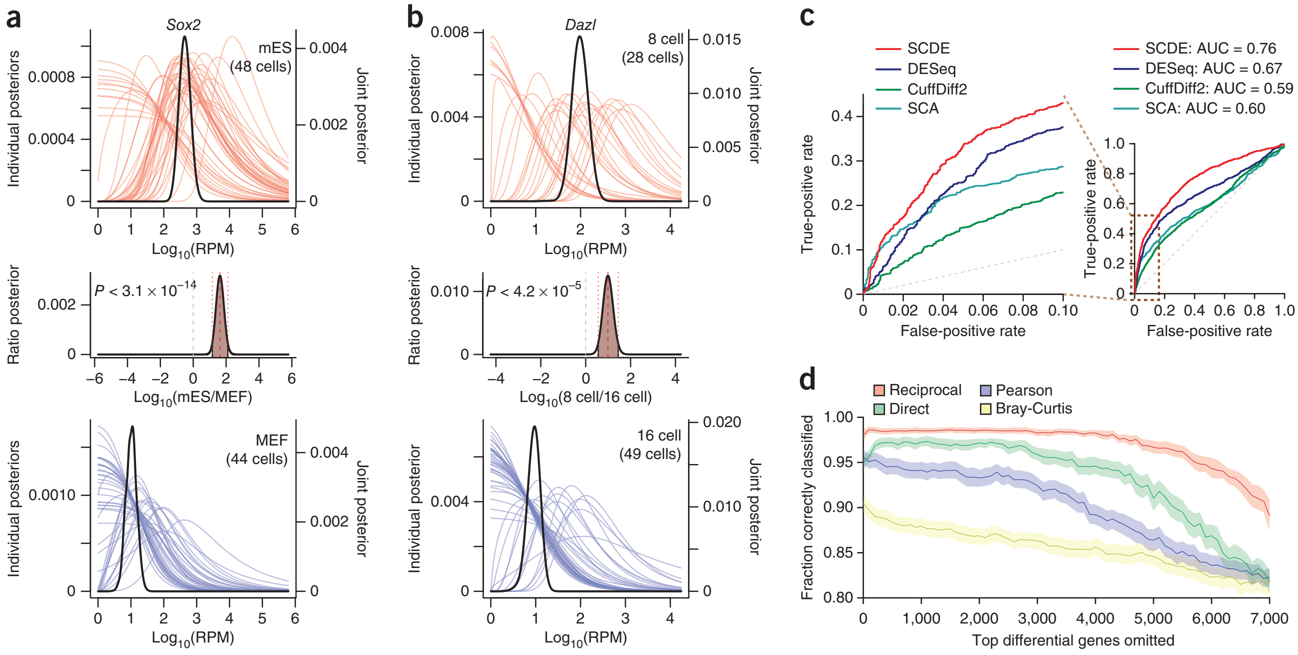

Figure 2 | Applying single-cell models for differential expression and subpopulation analyses. (**a**) Expression differences of *Sox2* between all ES and MEF cells, measured by Islam *et al.*[2]. The plots show posterior probability (*y* axis, probability density) of expression magnitudes in mouse ES (mES, top) and MEF (bottom) cells. The model fitted for each single cell is used to estimate the likelihood that a gene is expressed at any particular level, given the observed data (red or blue curves). The black curve shows the estimated joint posterior distribution for the overall level for each cell type. The posterior probability of the fold-expression difference is shown in the middle plot with the associated raw *P* value (two-sided) of differential expression. (**b**) Expression differences of *Dazl* between cells of 8-cell and 16-cell mouse embryo stages[12], as in **a**. A regulatory factor expressed in mammalian embryos[19,20], *Dazl* is expressed at earlier stages and shows a drop-off between 8- and 16-cell stages. (**c**) Receiver operating characteristic curves comparing the ability to detect differentially expressed genes, with bulk expression measurements as a benchmark[17]. SCA, single-cell assay[15]; AUC, area under curve. (**d**) Performance of error model–based transcriptional similarity measures in distinguishing ES and MEF cell types. The plot shows the fraction of correctly classified cells, assessed for increasingly difficult classification problems by iterative exclusion of up to 7,000 of the most informative genes (i.e., genes differentially expressed between ES and MEF, *x* axis). The 95% confidence bands (of the mean) are shown in light shading.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank X. Wang (Harvard Medical School) for help with packaging the implementation and F. Ferrari (Harvard Medical School) and M.B. Johnson (Children's Hospital, Boston) for critical review of the manuscript and SCDE implementation. This work was supported by US National Institutes of Health (NIH) grant K25AG037596 to P.V.K., fellowship awards from Leukemia and Lymphoma Research UK and Leukemia and Lymphoma Society to L.S. and NIH grants R01DK050234-15A1 and R01HL097794-03 to D.T.S.

AUTHOR CONTRIBUTIONS

P.V.K. conceived and implemented the computational approach. L.S. and D.T.S. designed and carried out the initial experimental study that led to the development of the presented approach.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
2. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
3. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Reports* **2**, 666–673 (2012).
4. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
5. Dalerba, P. *et al. Nat. Biotechnol.* **29**, 1120–1127 (2011).
6. Tang, F. *et al. PLoS ONE* **6**, e21208 (2011).
7. Brouilette, S. *et al. Dev. Dyn.* **241**, 1584–1590 (2012).
8. Buganim, Y. *et al. Cell* **150**, 1209–1222 (2012).
9. Munsky, B., Neuert, G. & van Oudenaarden, A. *Science* **336**, 183–187 (2012).
10. Brennecke, P. *et al. Nat. Methods* **10**, 1093–1095 (2013).
11. Wills, Q.F. *et al. Nat. Biotechnol.* **31**, 748–752 (2013).
12. Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. *Science* **343**, 193–196 (2014).
13. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
14. Trapnell, C. *et al. Nat. Biotechnol.* **31**, 46–53 (2013).
15. McDavid, A. *et al. Bioinformatics* **29**, 461–467 (2013).
16. Robinson, M.D. & Smyth, G.K. *Bioinformatics* **23**, 2881–2887 (2007).
17. Moliner, A., Enfors, P., Ibanez, C.F. & Andang, M. *Stem Cells Dev.* **17**, 233–243 (2008).
18. Tischler, J. & Surani, M.A. *Curr. Opin. Biotechnol.* **24**, 69–78 (2013).
19. Cauffman, G. *et al. Mol. Hum. Reprod.* **11**, 405–411 (2005).
20. Pan, H.A. *et al. Fertil. Steril.* **89**, 1324–1327 (2008).

ONLINE METHODS

Data sets and initial abundance estimates. ES and MEF single-cell measurements (96 cells) from Islam *et al.*[2] were used. The initial RPM estimates were obtained with TopHat[21] and HTSeq. The mouse embryo data were taken from Deng *et al.*, with the read alignments described in the manuscript[12].

Fitting individual error models. To identify a subset of genes that can be used to fit error models for particular single-cell measurements, all pairs of individual cells belonging to a given subpopulation (for example, all MEF cells) were analyzed with a three-component mixture model. To do so, the observed abundance of a given transcript in each cell was modeled as a mixture of the dropout (Poisson) and 'amplification' (negative binomial, NB) components. This way, the expression of a gene with observed RPM levels of r_1 and r_2 in cells c_1 and c_2, respectively, was modeled as

$$\begin{cases} r_1 \approx Poisson(\lambda_0) & \text{Dropout in } c_1 \\ \begin{cases} r_1 \approx NB(r_2) \\ r_2 \approx NB(r_1) \end{cases} & \text{Amplified} \\ r_2 \approx Poisson(\lambda_0) & \text{Dropout in } c_2 \end{cases}$$

The background read frequency for the dropout components was set at $\lambda_0 = 0.1$. The mixing between the three components was determined by a multinomial logistic regression on a mixing parameter $m = \log(r_1) + \log(r_2)$. Pseudocounts of 1 were added to r_1 and r_2 for log transformations. The mixture was fit with an EM algorithm, implemented under the FlexMix framework[22]. Alternatively, the initial three-component segmentation can be determined on the basis of a user-defined background threshold, which is much less computationally intensive. The genes that were assigned to the amplified components were noted, and a set of genes appearing in the amplified components in at least 20% of all pairwise comparisons of cells of the same subpopulation (excluding the cell for which the model was being fit) was used to fit the individual error models, as described below. The expected expression magnitude of these genes was estimated as a median observed magnitude between all the cell measurements in which a gene was classified to be in the amplified component. The aim of the 20% threshold is to have a sufficiently large number of measurements for a given gene so that the median expression-magnitude estimate would be reliable, and the model parameters resulting from the fitting procedure would correlate well for a range of values corresponding to 6–12 cells (**Supplementary Fig. 3d**).

To fit an individual error model Ω_c for a measurement of a single cell c, the observed RPM values were modeled as a function of an expected expression magnitude, with the set of estimates for a subset of genes described in the previous paragraph. The RPM level r_c observed for a gene in cell c was modeled as a mixture of a dropout and amplified components, as a function of an expected expression magnitude e, as

$$\begin{cases} r_c \approx NB(e) & \text{Amplified} \\ r_c \approx Poisson(\lambda_0) & \text{Dropout} \end{cases}$$

with the mixing parameter $m = \log(e)$. For each cell, the model Ω_c was fit with an EM algorithm based on the set of genes for which expected expression magnitudes have been obtained. The resulting estimates of parameters for the negative binomial and concomitant (mixing) regression were used as a description of an error model Ω_c in the subsequent analysis.

Differential expression analysis. With a Bayesian approach, the posterior probability of a gene being expressed at an average level x in a subpopulation of cells S was determined as an expected value (E) according to

$$p_S(x) = E\left[\prod_{c \in B} p(x \mid r_c, \Omega_c)\right]$$

where B is a bootstrap sample of S, and $p(x|r_c,\Omega_c)$ is the posterior probability for a given cell c, according to

$$p(x \mid r_c, \Omega_c) = p_d(x)p_{Poisson}(x) + (1 - p_d(x))p_{NB}(x \mid r_c)$$

where p_d is the probability of observing a dropout event in cell c for a gene expressed at an average level x in S, $p_{Poisson}(x)$ and $p_{NB}(x|r_c)$ are the probabilities of observing expression magnitude of r_c in case of a dropout (Poisson) or successful amplification (NB) of a gene expressed at level x in cell c, with the parameters of the distributions determined by the Ω_c fit. For the differential expression analysis, the posterior probability that the gene shows a fold expression difference of f between subpopulations S and G was evaluated as

$$p(f) = \sum_{x \in X} p_S(x)p_G(fx)$$

where x is the valid range of expression levels. The posterior distributions were renormalized to unity, and an empirical P value was determined to test for significance of expression difference.

Comparison of differential expression performance. The results of SCDE, DESeq, CuffDiff2 and single-cell assay (SCA) were benchmarked against an expression data set by Moliner *et al.*[17] that measured bulk MEF and ES cells grown with the same suspension growth protocol[23] as used by Islam *et al.*[2]. The ability to recover the top 1,000 genes showing the highest expression difference in Moliner *et al.* was assessed with ROC/AUC (**Fig. 2c** and **Supplementary Fig. 4**), ranking genes by significance of differential expression as determined by different methods.

Similarity measures and subpopulation analysis. The standard measure of the genome-wide similarity between two single-cell measurements was determined as a Pearson linear coefficient on log-transformed RPM values. Genes that did not show expression signals in any of the cells were excluded from the analysis. The Bray-Curtis similarity measure was also calculated on log-transformed values (and linear-based values showed lower performance).

The direct dropout similarity measure aims to estimate Pearson linear correlation excluding likely dropout events in any given cell. To achieve that, we evaluated average correlation across 1,000 sampling rounds, in each round probabilistically excluding likely dropout observations. Specifically, in each round, an observation of a given gene at an expression level x in a particular cell was substituted with a missing value with probability $p_d(x)k$, where $p_d(x)$ is the probability of a dropout event in the current cell at an expression-magnitude level x, and $k = 0.9$ is an additional factor (to stabilize similarity measure in cases when dropout rates are very high in a given cell). The overall similarity between any two cells was then calculated as an average (across 1,000 sampling rounds) Pearson linear correlation between log-transformed values of observations that are valid (not missing) in both cells.

The reciprocal dropout similarity measure aims to reduce the impact of dropout events on the Pearson linear correlation measure by weighting down the contribution of genes that are not likely to be reliably measured in both cells. For instance, if a gene was observed at a level x_1 in the first cell, we will weigh its contribution by the likelihood that such level of expression can be reliably detected (i.e., without dropout) in the second cell. This kind of reciprocal weighting minimizes the contribution of discrepant (i.e., amplified versus dropout) measurements to the overall similarity. Specifically, the reciprocal dropout similarity was calculated as a weighted Pearson linear correlation on log-transformed RPM values, weighting the contribution of each gene by

$$k\sqrt{(1-p_d^1(x_2))(1-p_d^2(x_1))}+(1-k)$$

where $p^1{}_d(x_2)$ is a probability of observing a dropout event in cell 1 for an expression magnitude x_2 at which the gene was observed in cell 2. $k = 0.95$ was used in calculating reciprocal dropout similarity. We find that both direct and reciprocal similarity measures show robust improvements in classification performance for a range of k values above 0.85 (**Supplementary Fig. 3e**).

All similarity measures do well when all 90+ cells and a complete gene set are considered. To provide a meaningful comparison, we measured performance on more challenging classification problems based on partial data. Specifically, a subset of 20 random ES and 20 MEF single-cell measurements was sampled in each iteration. Furthermore, an increasing fraction of top differentially expressed genes was excluded from the analysis (**Fig. 2d**, x axis) to pose a more challenging classification problem. The cells were clustered with the Ward method. The fraction of correctly classified cells was determined on the basis of the top-level split of the resulting clustering. The performance was evaluated on the basis of 200 such random sampling iterations.

Implementation. The algorithms were implemented as an R package, which is available for download at http://pklab.med.harvard.edu/scde/ or as **Supplementary Software**.

21. Trapnell, C., Pachter, L. & Salzberg, S.L. *Bioinformatics* **25**, 1105–1111 (2009).
22. Grün, B., Scharl, T. & Leisch, F. *Bioinformatics* **28**, 222–228 (2012).
23. Andäng, M., Moliner, A., Doege, C.A., Ibanez, C.F. & Ernfors, P. *Nat. Protoc.* **3**, 1013–1017 (2008).

doi:10.1038/nmeth.2967