DEsingle: A new method for single-cell differentially expressed genes detection and classification

Zhun Miao[1], Xuegong Zhang[1, 2, *]

[1] MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China
[2] School of Life Sciences, Tsinghua University, Beijing 100084, China

* To whom correspondence should be addressed. Tel: +86-10-62794919; Fax: 86-10-62773552; Email: zhangxg@tsinghua.edu.cn

ABSTRACT

There are excessive zero values in single-cell RNA-seq (scRNA-seq) data. Some of them are real zeros of non-expressed genes, while the others are the so-called “dropout” zeros caused by the low mRNA capture efficiency of tiny amounts of mRNAs in single cells. These two types of zeros should be distinguished in differential expression (DE) analysis and other types of analyses of scRNA-seq data. We proposed a new method DEsingle for DE analysis in scRNA-seq data by employing the Zero-Inflated Negative Binomial (ZINB) model. We proved that DEsingle could estimate the percentage of real zeros and dropout zeros by modelling the mRNA capture procedure. According to this model, DEsingle can distinguish three types of differential expression between two groups of single cells, with regard to differences in expression status, in expression abundances, and in both. We validated the performance of the method on simulation data and applied it on real scRNA-seq data of human preimplantation embryonic cells of different days of embryo development. Results showed that DEsingle outperforms existing methods for scRNA-seq DE analysis, and can reveal different types of DE genes that are enriched in different functions.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a new technology developed in recent years which can study transcriptomes in individual cells (1-4). The expression level of a gene in a cell could be estimated by counting the number of sequenced reads mapped to the reference sequence of the gene (5-7). The importance of measuring the gene expression level in single cells as well as the importance of studying bioinformatics methods for single-cell data has been increasingly recognized (7-10).

Differential expression (DE) analysis is to detect genes whose expression levels are significantly different between the compared groups of samples (11-15). It has been a key task in transcriptome study since the early days of microarrays. Traditional DE analysis methods for RNA-seq data were designed for bulk RNA sequencing (11,16,17), which needs millions of cells in one sample (6,18), and

most of those DE analysis methods focus on the detection of DE genes by their mean expression levels (11,12,16,17,19).

scRNA-seq data has many different characteristics from bulk RNA-seq data (20). One important difference is that there are much more zero values in scRNA-seq data than in bulk RNA-seq data (12). Due to the tiny amount of mRNAs in one cell (~0.01-2.5pg), the small mRNA copy number of each gene in a cell (thousands of genes have only 1-30 mRNA copies) and the very low mRNA capture efficiency (~5-25%, up to ~40%) (21-24), some mRNAs are totally missed during the reverse transcription step and the following cDNA amplification step, and consequently undetectable in the later sequencing step (3). This phenomenon is called dropout events (25,26). We call this type of zero values as dropout zeros. On the other hand, because of the heterogeneity between cells and the stochastic nature of transcription in a single cell, there is also a high chance that the expression level of some genes are really zero when it is sequenced (19,27). In other words, these genes are not expressed in the cell when the cell is lysed. We call this type of zero values as real zeros. The excessive zero values observed in most scRNA-seq data are mixed with these two possible types of zeros.

The transcription process of a gene in one single cell is an on-off stochastic process (28,29). If there is not a single copy of the mRNA molecule of a gene in the cell at the time of RNA capturing, scRNA-seq will produce a zero count for this gene. This zero is a real zero. It reflects the expression status of the gene in the cell. For a group of cells, the proportion of cells with real zeros reflects the overall expression status of the gene in the group (19). Existing methods for analysing differential gene expression of single cell did not take this into consideration.

In this paper, we proposed a new method called DEsingle for DE analysis of scRNA-seq data. DEsingle is based on Zero-Inflated Negative Binomial (ZINB) model (30-33), which provides a good fit for the data. It introduced an extra parameter to model the excessive amount of zeros in the data. Using this model, we can separate observed zero values as two parts to roughly reflect the proportion of real zeros and dropout zeros due to very low expression level. Simulation experiments showed that DEsingle can not only detect DE genes with better performance than existing methods, but can also distinguish different expression status of a gene from differential expression abundance of the gene. It can report three types of differential expression: different expression status (DEs), differential expression abundance (DEa) and general differential expression (DEg) with mixed effects of both DEs and DEa. Among the three types, DEa and DEg are similar to conventional understanding on differential expression, but the type DEs highlights a new situation that has not been paid sufficient attention. We compared our method with several representative existing methods on simulation data. The methods we compared with include traditional DE detection methods edgeR (16) and DEGseq (17) that were developed for bulk RNA-seq data but have been also used on many scRNA-seq data (12), as well as new methods specifically developed for scRNA-seq data including BPSC (34), D3E (19), monocle (35), SCDE (25). We applied DEsingle on a real scRNA-seq dataset of human preimplantation embryonic cells of different days of embryo development (36) and found interesting

observations on the three types of differentially expressed genes between different days of the development.

MATERIAL AND METHODS

Using the ZINB model to fit scRNA-seq data with excessive zeros

The zero percentages are very high for most of genes in scRNA-seq data. Figure 1A shows the histogram of zero percentages of all the expressed genes in a human embryonic scRNA-seq dataset (36). Because of this excessive proportion of zeros as shown in the data, we consider using zero-inflated models in our method. For RNA-seq read counts data, Negative Binomial (NB) distribution has been widely used in most DE analysis methods (e.g., edgeR (16), DESeq (11), baySeq (37)). We adopted the Zero-Inflated Negative Binomial (ZINB) model in our method and found that it can fit scRNA-seq data well (Figure 1B).

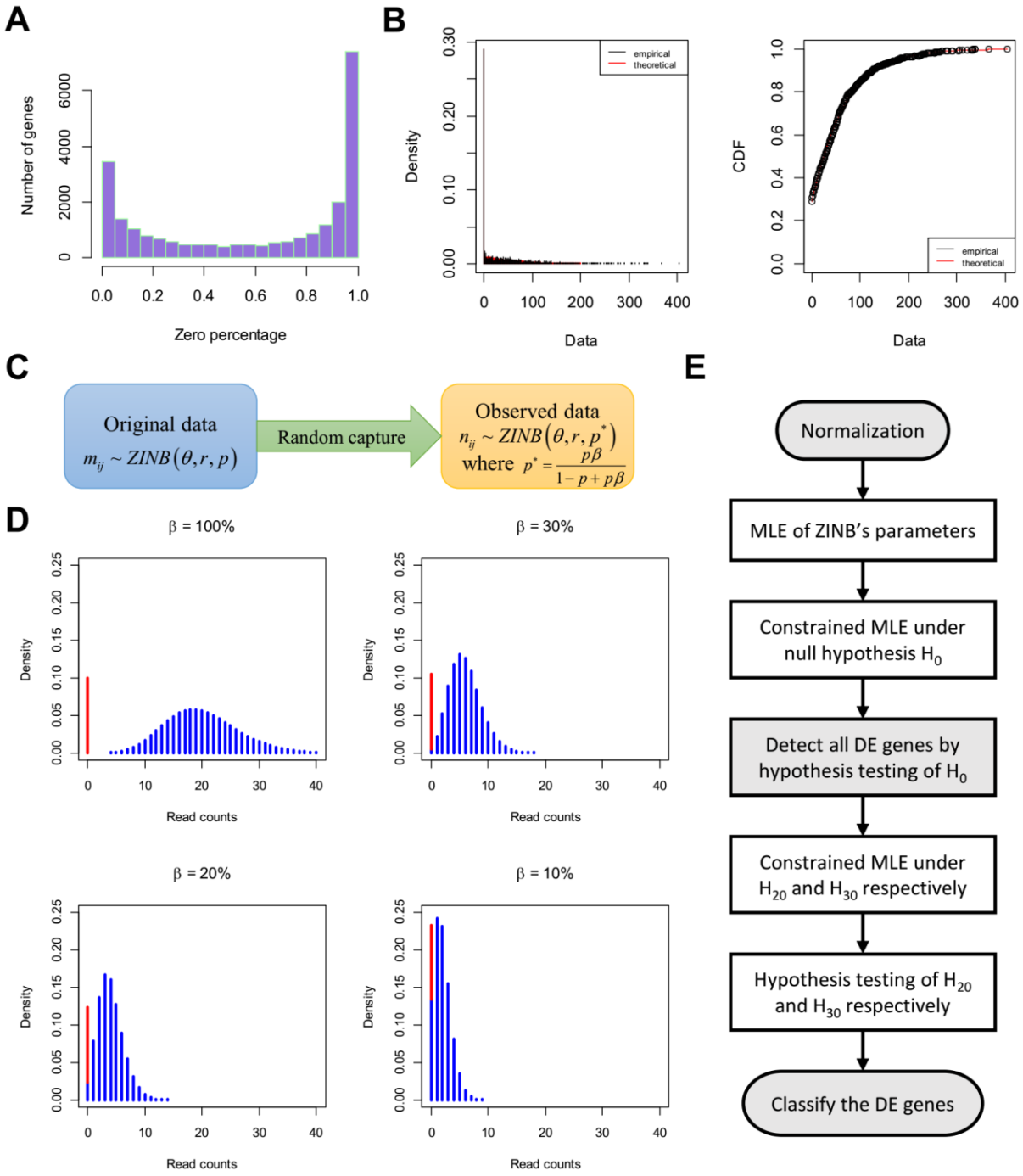

Figure 1. ZINB model for scRNA-seq data and workflow of DEsingle. (A) Histogram of zero percentages of all expressed genes in a human embryonic scRNA-seq dataset. (B) An example of ZINB model fitting for scRNA-seq data. The figure is density fitting (left) and cumulative distribution function fitting (right) for the scRNA-seq data (empirical) to the ZINB model (theoretical). (C) Mathematical modeling of mRNA capture procedure, under the random capture assumption. (D) Theoretical ZINB distribution of a gene with different random capture efficiency β denoted above each graph. β = 100% represents the distribution of original data; β = 30%, 20% and 10% represent the observed data obtained from the original data after mRNA capture procedure. The parameters of ZINB distribution of the original data are θ = 0.1, r = 20 and p = 0.5. The red line represents the probability density of real zero expression (constant zeros), which is comes from the θ parameter of the ZINB model; the blue line represents the probability density of NB part of ZINB model. When β becomes

smaller, the zero density from NB part (blue line on zero value) becomes larger. (E) Workflow of DEsingle to detect and classify DE genes. Hypothesis testing of $H_0 : \theta_1 = \theta_2, r_1 = r_2, p_1 = p_2$ is used to detect all the DE genes; hypothesis testing of $H_{20} : \theta_1 = \theta_2$ and $H_{30} : r_1 = r_2, p_1 = p_2$ are used to classify the found DE genes.

The ZINB model is a mixture of constant zeros and NB distribution, with mixture proportions of θ and $1-\theta$, respectively. The probability mass function (pmf) of ZINB distribution for the read counts N_g of gene g in a group of cells is

$$P\left(N_g = n \mid \theta, r, p\right) = \theta \cdot I\left(n = 0\right) + \left(1-\theta\right) \cdot NB\left(r, p\right), \quad n = 0,1,2,...$$

where N_g is the read counts of gene g, θ is the proportion of constant zeros of gene g in the group of cells, r is the size parameter and p is the probability parameter of the NB distribution part of the ZINB model. Note that the NB part can also have zero values. The observed zero values are the sum of constant zeros and the zero values from the NB distribution part.

Model the mRNA capture procedure

Most dropout zeros are produced because of the very low mRNA capture efficiency (~5-25%, up to ~40%) and the small mRNA copy number of each gene (thousands of genes only have 1-30 mRNA copies) in a cell (21-24). As a result, mRNA molecules in a cell can be randomly missed during the reverse transcription step and the following cDNA amplification step, and the mRNA products of some genes may be totally missed in the capturing procedure, which then produces dropout zeros in the scRNA-seq data (3,25,26). In this section, we try to model this mRNA capture procedure and study what impact this process will have on the ZINB distribution. For convenience, we use “mRNA capture procedure” to refer to both the mRNA capture procedure in reverse transcription step and the cDNA capture procedure in cDNA amplification step.

Let m_{ij} denote the original transcript copy number of gene i in cell j, and let n_{ij} denote the number of captured transcript copies of gene i in cell j. In real data, the true m_{ij} is unknown, and the observed n_{ij} is the result of sampling from the existing transcripts with the mRNA capture procedure. In practice, the cDNA amplification step and sequencing from the cDNA library can also introduce noises and bias in the number of reads sequenced from captured transcripts. But this is less severe comparing to the mRNA capture procedure. Therefore we assumed that the estimated abundance of the transcript of gene i in cell j from the sequencing reads reflects the n_{ij}.

We use a parameter $\beta \in \left[0\%, 100\%\right]$ to denote the efficiency of the mRNA capture procedure. Assume $m_{ij} \sim ZINB\left(\theta, r, p\right)$, under the random capture assumption that all transcripts are captured

with the same probability β, we can prove that $n_{ij} \sim ZINB\left(\theta, r, p^*\right)$ with $p^* = \frac{p\beta}{1-p+p\beta} \leq p$ (Figure 1C), as described in Supplementary Data.

Therefore, we can see that,

1. Parameter p in the ZINB model becomes $p^* = \frac{p\beta}{1-p+p\beta} \leq p$ after the capture procedure.

   Since the mean and the variance of the NB part of the ZINB model are $\frac{pr}{1-p}$ and $\frac{pr}{(1-p)^2}$ respectively, when p becomes smaller, both the mean and the variance of the NB part become smaller. The NB distribution moves to closer toward zero after the capture procedure. This means more zeros in the observed data can be from the NB part of the model. Figure 1D shows examples of theoretical ZINB distribution of original data (β = 100%) and observed data with different capture efficiency β (30%, 20%, 10%).

2. Parameter θ is unchanged after the mRNA capture procedure, which means that the proportion of constant zeros of the gene among the group of cells is unchanged. For the ideal original data (β = 100%), only those cells with no transcript of the gene in the cell will give zero values. Those are real zeros that reflect genes that are not at the transcription status. For any none-100% capture efficiency, observed zeros are the mixture of real zeros and dropout zeros. But as θ is unchanged by the capture procedure, the θ estimated from observed data n_{ij} can be used to measure the proportion of real zeros that represent the expression status of the gene in the group of cells.

Figure 1E is a workflow of DEsingle. It includes three major steps: the normalization of the data, detection of the DE genes and classification of the found DE genes into subtypes: DEs, DEa or DEg. The input of DEsingle is the raw read counts matrix from scRNA-seq data. DEsingle integrates a median normalization method proposed by DESeq to normalize the data (11), as described in Supplementary Data. After normalization, maximum likelihood estimation (MLE) and constrained MLE of the parameters of two ZINB populations are calculated. Finally, likelihood ratio tests are conducted to detect the DE genes and classify them into three types.

Comparing two ZINB populations

Detecting differentially expressed genes of two groups of cells using ZINB model is equivalent to testing the heterogeneity of two ZINB populations (30-32,38). When any of the three parameters of two ZINB models has significant difference, we consider the gene is a DE gene. We use $\left\{n_{ij_1}\right\}$ and $\left\{n_{ij_2}\right\}$ to denote the read counts of gene i in cell j of group 1 and of group 2, respectively, and use the following 3 steps to test the difference of the two populations.

1. Calculate MLE of $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{r}_1, \hat{r}_2, \hat{p}_1, \hat{p}_2\right)$ for the parameters of the two ZINB populations using Expectation-Maximization (EM) algorithm (39), i.e.,

$$\left\{n_{ij_1}\right\} \xrightarrow{MLE} \hat{\boldsymbol{\theta}}_1 = \left(\hat{\theta}_1, \hat{r}_1, \hat{p}_1\right), \ \left\{n_{ij_2}\right\} \xrightarrow{MLE} \hat{\boldsymbol{\theta}}_2 = \left(\hat{\theta}_2, \hat{r}_2, \hat{p}_2\right)$$

2. Calculate constrained MLE of $\hat{\boldsymbol{\theta}}_0 = \left(\hat{\theta}_{1,0}, \hat{\theta}_{2,0}, \hat{r}_{1,0}, \hat{r}_{2,0}, \hat{p}_{1,0}, \hat{p}_{2,0}\right)$ under the null hypothesis $H_0 : \theta_1 = \theta_2, r_1 = r_2, p_1 = p_2$, i.e.,

$$\left\{n_{ij_1}\right\} \xrightarrow{constrained\ MLE} \hat{\boldsymbol{\theta}}_{1,0} = \left(\hat{\theta}_{1,0}, \hat{r}_{1,0}, \hat{p}_{1,0}\right)$$

$$\left\{n_{ij_2}\right\} \xrightarrow{constrained\ MLE} \hat{\boldsymbol{\theta}}_{2,0} = \left(\hat{\theta}_{2,0}, \hat{r}_{2,0}, \hat{p}_{2,0}\right)$$

They are equivalent to calculate unconstrained MLE using the two groups of counts data together,

$$\left\{n_{ij_1}, n_{ij_2}\right\} \xrightarrow{MLE} \hat{\boldsymbol{\theta}}_0 = \left(\hat{\theta}_{1,0} = \hat{\theta}_{2,0}, \hat{r}_{1,0} = \hat{r}_{2,0}, \hat{p}_{1,0} = \hat{p}_{2,0}\right).$$

3. Hypothesis testing. According to Wilks Theorem (40), under the null hypothesis H_0, the χ^2_{LR1} statistics follows a χ^2_3 distribution,

$$\chi^2_{LR1} = -2\log\lambda(y) = -2\log\frac{\sup_{\Theta_0} L(\theta \mid y)}{\sup_{\Theta} L(\theta \mid y)} = -2\left[l\left(\hat{\boldsymbol{\theta}}_0\right) - l\left(\hat{\boldsymbol{\theta}}\right)\right] = 2\left[l\left(\hat{\boldsymbol{\theta}}\right) - l\left(\hat{\boldsymbol{\theta}}_0\right)\right] \sim \chi^2_3$$

Then hypothesis testing is conducted using the χ^2_{LR1} statistics.

Detailed algorithms of the parameter estimation and statistical test are provided in the Supplementary Data.

Three types of DE genes

The three types of DE genes between single cell groups can be identified according to patterns of differences of parameters in the following tests. The testing against the null hypothesis $H_0 : \theta_1 = \theta_2, r_1 = r_2, p_1 = p_2$ detects all genes that are significantly differentially expressed between the two groups in any of the parameters. Then we'll use another two hypothesis tests $H_{20} : \theta_1 = \theta_2$ and $H_{30} : r_1 = r_2, p_1 = p_2$ to classify the found DE genes into three types or categories.

The first type is of genes with H_0 and H_{20} rejected but with H_{30} accepted. This means that there is a significant difference in the number of zero values of the gene between the two groups of cells, but the difference between the cells with non-zero expression values of the gene shows no significance. We call this type of DE genes as DEs genes, meaning that they have different expression statuses between the two groups. The second type is of genes with H_0 and H_{30} rejected but H_{20} accepted. We call them DEa genes as they have differential expression abundances in the cells with non-zero values, which the relative proportion of zero values do not show significance difference between the two groups of cells. The third type is of genes that have both different expression statuses in the two

groups and differential expression abundances. Genes for which H_0 are rejected but H_{20} and H_{30} are either both accepted or both rejected are of this category. We call these genes as DEg genes, meaning that they have general differential expression between the two groups of cells.

To analyze the direction of the differential expression, we further breakdown the 3 types into 6 sub-categories. A DEs gene is more active in the group with significantly less zero-valued cells than the other group. In other words, the gene is in "turned-on" status in more cells of this group while the other group has more cells with this gene "turned-off". We call that the gene is "DEs-on" in the group with less zeros and is "DEs-off" in the group with more zeros for convenience. For a DEa gene, the proportions of zeros are not significantly different between the two groups of cells, but for those none-zero parts, the gene is significantly highly expressed in one group than in the other group. We call the gene as "DEa-up" in the highly expressed group and "DEa-down" in the lowly expressed group. Similarly, for a DEg gene, there are a significant expression difference between the two groups of cells, but the difference is not solely caused by proportion of zeros or by expression of the non-zero values. When a DEg gene has less zeros and also has higher expression values in one group than the other group, we call the gene as "DEg-up" in this group, and call it as "DEg-down" in the other group. It can be imagined that there could be cases that a gene has more zeros in one group but has higher expression for the non-zero values in this group, or that having less zeros is accompanied by lower expression of non-zero values. In such cases, it will not be feasible to use a single "up" or "down" to describe the direction of the difference. However, we observed only few such cases in real data so we do not count for such special cases when profiling a whole dataset. Table 1 summarized the 3 types of DE genes and the 6 sub-categories.

Table 1. Sub-categories of DE genes. $\mu_{NB} = \frac{pr}{1-p}$, which is the mean of the NB part of the ZINB model.

DE type	DE parameter	Null Hypotheses Rejection			Conditions for "DEs-on/DEa-up/DEg-up" in group 1
		H_0	H_{20}	H_{30}	
Different Expression Status (DEs)	θ	Significant	Significant	Not significant	$\theta_1 \le \theta_2$
Different Expression Abundance (DEa)	r or p	Significant	Not significant	Significant	$\mu_{NB_1} \ge \mu_{NB_2}$
General Differential Expression (DEg)	θ, and r or p	Significant	Both significant or both not significant		$\mu_{NB_1}(1-\theta_1) \ge \mu_{NB_2}(1-\theta_2)$

Simulation data

We generated a series of simulation data to study the performance of the proposed method. To generate scRNA-seq simulation data more fairly, we used the simulation method proposed by (19,26). In brief, the simulated scRNA-seq data was generated by randomly sampling from a Poisson-Beta distribution (19), which is originated from the transcriptional bursting model (41). The data was produced with the following procedures: Firstly, draw a variable c from a Beta distribution with

parameters α and β, namely $c \sim Beta(\alpha, \beta)$. Secondly, randomly sample a number from the Poisson distribution with parameter $\lambda = c\gamma$. Parameters α, β and γ is randomly picked from the parameters list downloaded from the additional file of (19), which was generated from a real scRNA-seq dataset (42).

To add dropout events to the simulated data, we applied the dropout model introduced by Pierson and Yau (26). Specifically, let x_{ij} denotes the expression level of gene i in cell j, and let μ denote the mean of non-zero expression level (log read count) of gene i across cells. The dropout rate of gene i is modeled as $p_0 = \exp(-\lambda\mu^2)$, where λ is the exponential decay parameter and is shared across genes. We use $\lambda = 0.1$ in our simulation as recommended. Then the expression level y_{ij} of gene i in cell j after adding dropout events is denoted as $y_{ij} = \begin{cases} x_{ij}, & \text{if } h_{ij} = 0, \\ 0, & \text{if } h_{ij} = 1, \end{cases}$, where $h_{ij} \mid x_{ij} \sim \text{Bernoulli}(p_0)$.

Using this strategy, we simulated two types of scRNA-seq data: those with and without additional dropout events. For each type of data, we simulated datasets with 10×2, 50×2, 100×2 and 200×2 cells, each with 10,000 simulated genes. We got 8 sets of simulation data in this way, 4 with dropout events and 4 without. For each dataset, we set half of the genes as DE genes. A non-DE gene is defined as a gene all three parameters have changes of less than 1.5 fold between the two groups. And a gene with any of the three parameters showing fold change greater than 1.5 is regarded as a DE gene.

We applied the proposed DEsingle and 6 other methods BPSC (34), D3E (19), monocle (35), SCDE (25), edgeR (16), DEGseq (17) on the 8 sets of simulation data. Four of the methods, BPSC, D3E, monocle, SCDE, are DE analysis methods specially designed for scRNA-seq. The other two methods, edgeR and DEGseq, are traditional DE analysis methods developed for bulk RNA-seq data. But they have also been widely applied on single-cell data. The parameters settings of each method in the experiments were provided in the Supplementary Data.

Real scRNA-seq data of human embryonic cells

We applied the DEsingle method on a public scRNA-seq dataset of human preimplantation embryonic cells (36). We conducted a systematic comparison on the gene expression of the 81 cells from embryonic day 3 (E3) and that of the 190 cells from embryonic day 4 (E4) in this dataset. We used the mapped raw read counts table provided by the original authors as input to DEsingle to detect and analyze the genes that are differentially expressed between E3 and E4 cells.

RESULTS

Performances on the simulation data

We compared DEsingle with the 6 existing methods on the simulation data. Figure 2 shows the Receiver Operating Characteristic (ROC) curve of the methods on the 8 simulation datasets. We used the Area Under Curve (AUC) to quantitatively evaluate the performance of the methods (43). We can see that DEsingle is almost always the best among the compared methods in the experiments, except for the comparison of 10 vs 10 cells without additional dropout events (upper left panel in Figure 2), in which DEsingle is the second best.

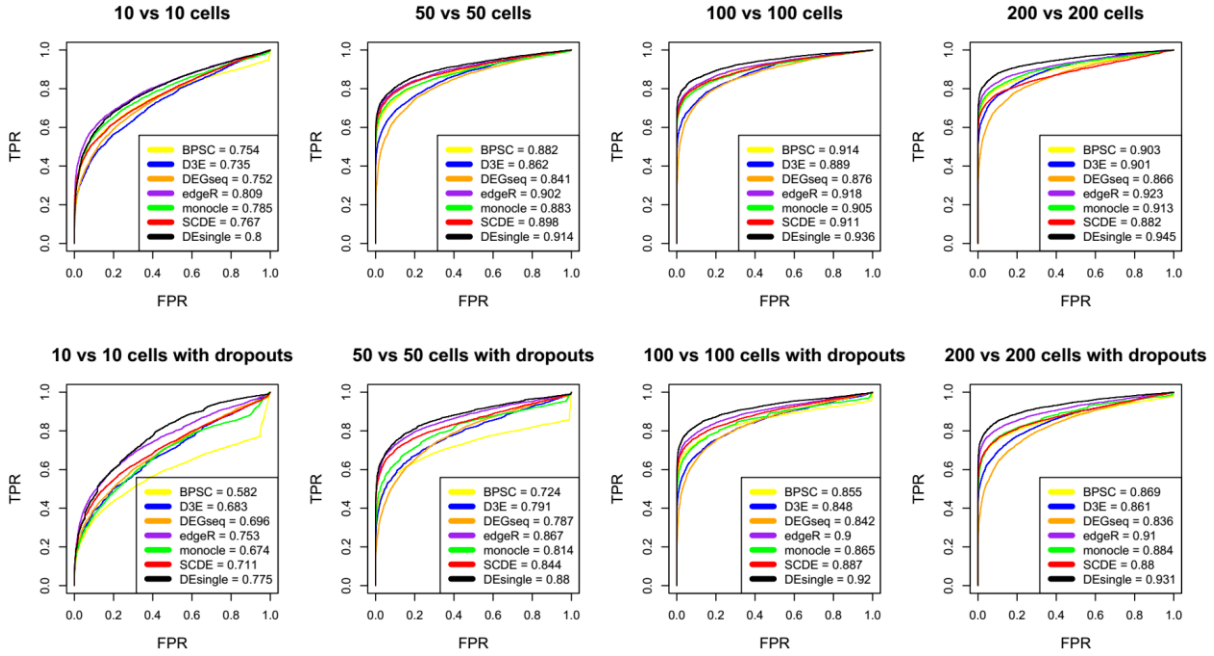

Figure 2. Evaluating performances of DE analysis methods on simulation data. ROC curves and AUCs of DE analysis using 7 methods on 8 scRNA-seq simulation datasets. The top 4 graphs are using simulation data without additional dropout events while the bottom 4 graphs are using simulation data with dropout events. The sample size for comparison is annotated above each graph. The AUC of each method is annotated on bottom right corner of each graph. We could see that, DEsingle performs best among the methods in these experiments except for the 10 vs 10 comparison without additional dropout events (the graph in upper left corner), in which DEsingle is the second best.

Three types of differentially expressed genes between E3 and E4 cells

When comparing the 81 cells of E3 to the 190 cells of E4 from the human preimplantation embryonic dataset (36), DEsingle reported a total of 7,560 genes that are differentially expressed at the level of p-value < 0.05 (Bonferroni correction). The majority of them, namely 5,685 genes (75.2%), belong to the DEg type. The differences on the proportion of zeros and on the mean expression levels of the non-zero values are of the same direction for most genes. Among them, 2,909 genes are DEg-up and 2,776 genes are DEg-down in E3 cells comparing to E4 cells. There are another 1,147 genes (15.2%) that are of the DEa type, with 670 DEa-up genes and 477 DEa-down genes in E3 comparing to E4. Besides these genes of the more conventional sense of differential expression, DEsingle reported 728 DEs genes (9.6%) with significant differences in the proportion of zero values in the two groups of

cells. Among them, 333 genes are DEs-on in E3 and 395 genes are DEs-off in E3 comparing to in E4. The E3 DEs-on genes are active in E3 but not active in E4, and the E3 DEs-off genes are not active in E3 but active in E4.

Figure 3A shows the heatmap of top 500 genes of each of the 3 types of DE genes. DE genes of the same type tends to be clustered together, and each type has its distinct expression pattern in E3 and E4. Figure 3B shows histograms of gene expression of 3 example genes in E3 cells and E4 cells. We can see that all three genes have different distributions among E3 cells and E4 cells, but the differences of different patterns. We are interested in whether the different patterns of DE imply different biological functions.

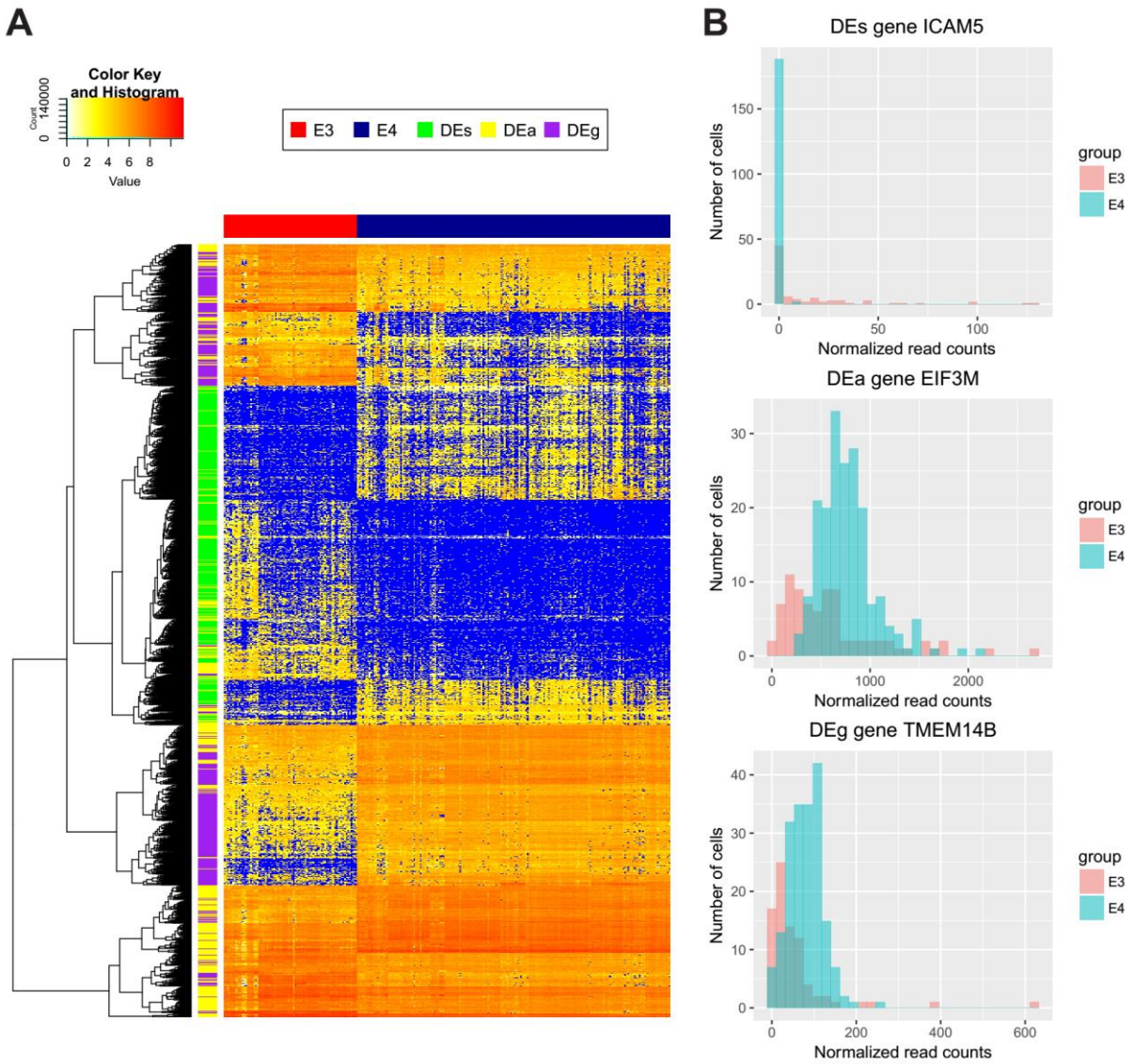

Figure 3. Heatmap and histogram of DE genes found by DEsingle. (A) log(#{counts} + 1) of top 500 of each of the 3 types DE genes of E3 versus E4 human preimplantation embryonic cells found by DEsingle. Row is gene and column is cell. Specially, for #{counts} = 0, the color of heatmap is set to blue to show the zero expression separately. We could see that same type of DE gene is basically clustered together and different types of DE genes show different patterns in the heatmap. (B) Histogram of expression levels of three DE genes examples

detected by DEsingle. From top to bottom, the three DE gene belongs to DEs, DEa and DEg category respectively. The sample number of E3 and E4 cell is 81 and 190 respectively.

To explore the underlying functional roles of each kind of DE genes, we performed Gene Ontology (GO) enrichment analysis on the top 500 of each of the 3 types of DE genes using DAVID (44). The 1,500 DE genes are divided into 6 sub-categories for GO enrichment analysis: 238 DEs-on genes, 262 DEs-off genes, 274 DEa-up genes, 226 DEa-down genes, 232 DEg-up genes and 268 DEg-down genes. The “on”/“off” and “up”/“down” here all refer to E3 comparing to E4. Table S1 listed all enriched GO terms with some highlighted in Figure 4B. Figure 4A shows the overlap between the enriched GO terms from the 6 sub-categories of DE genes. We can see that there are more enriched GO terms in DEa and DEg genes than in DEs genes, which implies that DEs genes are involved in less biological processes than DEa genes and DEg genes. However, the GO terms enriched in DEs genes have no intersection with the GO terms enriched by other DE genes (Figure 4A). This indicates that the regulation and function of genes that have different expression statuses between E3 and E4 involve different pathways with those of the genes that are differentially expressed in the conventional sense.

Specifically, DEs-on genes of E3 are enriched for Biological Process (BP) GO terms of cell adhesion, extracellular matrix organization, Cellular Component (CC) GO terms of integral component of plasma membrane, plasma membrane, basement membrane, cell junction, and Molecular Function (MF) GO terms of calcium ion binding, extracellular ligand-gated ion channel activity (Figure 4B; Table S1). E3 is around the 8-cell stage of preimplantation embryos (45), during which cell compaction occurs and functional gap junctions are formed (46,47). Cell compaction is initiated by the E-cadherin mediated cell adhesion (47-50), and the genes related to cytoskeletal, cell junction and cell adhesion are also involved in this process (51). The enriched GO terms of the DEs-on genes of E3 are all directly or indirectly associated with the cell compaction process. For example, GO terms calcium ion binding and extracellular ligand-gated ion channel activity are known to play an important role to the function and interactions of E-cadherin (52); GO terms extracellular matrix organization, integral component of plasma membrane, plasma membrane, basement membrane and cell junction are essential to the formation of gap junctions and cell adhesion (46,47,49,51). For example, the DEs E3-on genes ICAM1 (intercellular adhesion molecule 1) and ICAM5 (intercellular adhesion molecule 5) show up in 4 of the GO terms listed above. The zero expression ratio of gene ICAM1 and ICAM5 in E3 cells are 28.4% and 45.3% respectively, while the ratio in E4 cells are 73.6% and 98.2% respectively (Figure 3B). This is to say, these two genes are expressed in most of cells in E3 but are turned off in almost all cells in E4. Genes related with cell compaction are active in E3 but are shut down gradually after compaction is finished and morula is formed in E4 (Figure 4B). Recent study has shown that, although the initiation of cell compaction occurred from 4-cell to 16-cell stage (about late embryonic day 2 to embryonic day 4), most human embryos (86.1%) initiated compaction at the 8-cell stage or later (around E3 to E4 stage), with initiation at the 8-cell stage (E3) being most frequent (53). That explains why the compaction related genes are still expressed in some (very few) E4 cells.

DEs-on genes of E4 are enriched for the BP GO term of positive regulation of collagen biosynthetic process (Figure 4B; Table S1). According to the recent study of Petropoulos et al. (36), human preimplantation embryonic cells will differentiate into three lineages of epiblast (EPI), primitive endoderm (PE) and trophectoderm (TE) simultaneously at E5. And Rasmussen et al. showed that collagen could improve the differentiation of human embryonic cells towards endoderm (54), which implies some inner cells of morula will synthesize collagen at E4 to induce themselves to differentiate into PE cells at E5. That explains why the GO term of positive regulation of collagen biosynthetic process is enriched in DEs-on genes of E4 and also verifies the different expression status of the related genes between E3 and E4 cells.

For the other types of DE genes, DEa-up genes of E3 are enriched for BP GO terms of mRNA splicing via spliceosome, mRNA 3'-end processing, RNA splicing, mRNA export from nucleus, RNA export from nucleus, mRNA processing, and CC GO terms of nucleoplasm, nucleus, nucleolus (Figure 4B; Table S1). DEg-up genes of E3 also are enriched in CC GO terms of nucleus and nucleoplasm. These GO terms are consistent with the process of zygotic genome activation (ZGA) or embryonic genome activation (EGA) that occurs at E3 (36,45), which brings the burst of transcription (55) in nucleus (56,57) simultaneously (Figure 4B).

DEa-up genes of E4 are enriched in BP GO terms of translational initiation, translation, cytoplasmic translation, CC GO terms of ribosome, cytosol, cytoplasm, and MF GO terms of translation initiation factor activity (Figure 4B; Table S1). These enriched GO terms are in agreement with the fact that translation occurs on the ribosomes in the cytoplasm (56,57), and the translation peak of EGA is delayed (58,59) to E4, which is called uncoupling of transcription and translation during EGA (60).

Except for the GO terms related to translation, DEg-up genes of E4 are enriched for BP GO terms of mitochondrial translational elongation, mitochondrial respiratory chain complex I assembly, mitochondrial electron transport NADH to ubiquinone, mitochondrial translation, mitochondrial ATP synthesis coupled proton transport, protein targeting to mitochondrion, and CC GO terms of mitochondrial inner membrane, mitochondrion, mitochondrial respiratory chain complex I, and mitochondrial ribosome (Figure 4B; Table S1). All these enriched GO terms are directly or indirectly related to mitochondrion. Dumollard et al. have reported that before the compaction is complete, pyruvate is metabolized by mitochondria whereas glucose is not, and that after compaction, glucose becomes the major substrate for energy supply instead of pyruvate in embryonic cells (61) (Figure 4B). Moreover, Wilding et al. have showed that the aerobic respiration is upregulated and the percentage of glucose metabolised through aerobic respiration rises dramatically during this period (62). These reports as well as our observations imply that the major metabolic mode of mitochondrion is changed from E3 to E4 and the mitochondrion becomes more active at E4. It is reasonable to infer that the changes of the metabolic mode and activity of mitochondrion may be related to the translation peak in E4 which requires more energy than E3.

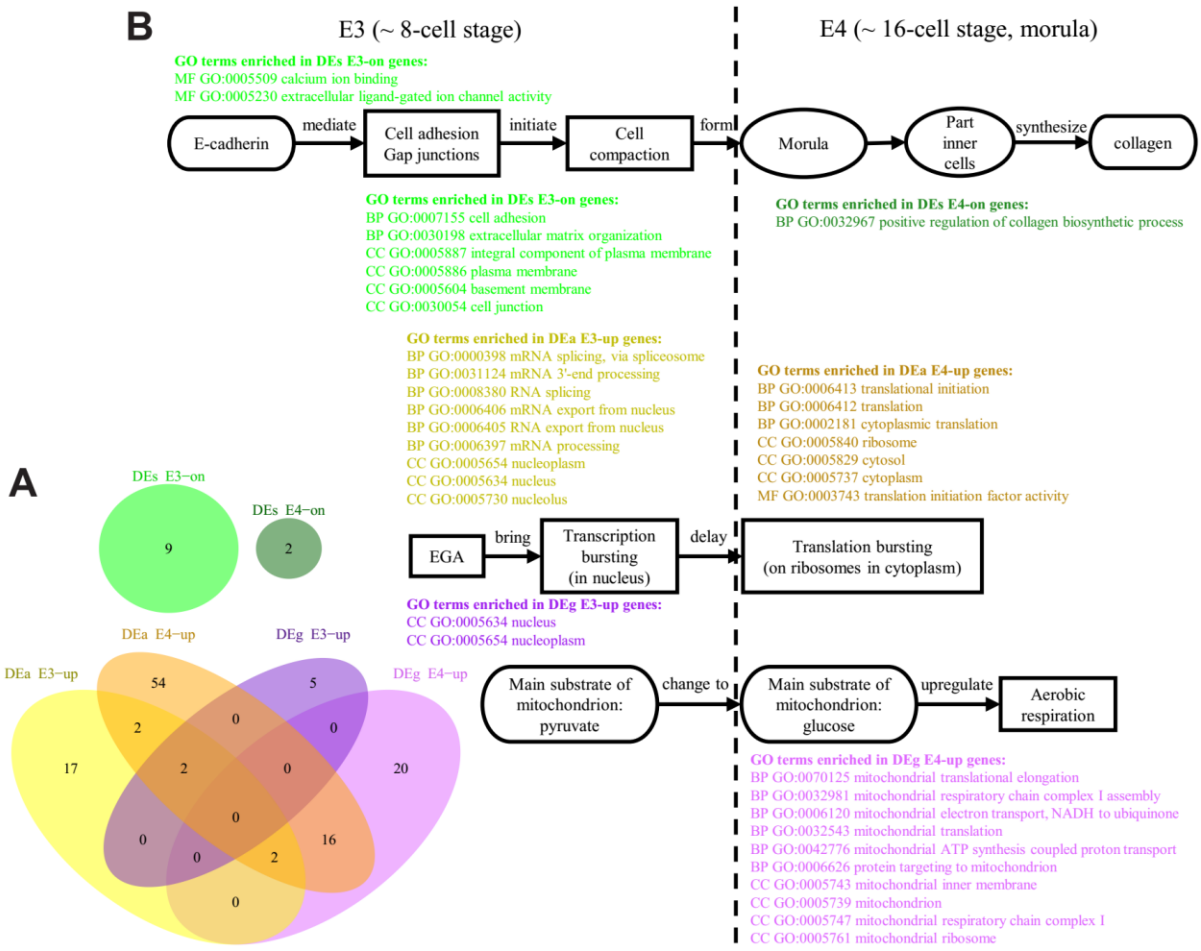

Figure 4. Venn diagram and related biological process of enriched GO terms in 6 sub-categories of DE genes. (A) Venn diagram of the enriched GO terms in the DE genes showed in Figure 3A. These 1500 DE genes are divided into 6 sub-categories for GO enrichment analysis: 238 DEs E3-on genes, 262 DEs E4-on genes, 274 DEa E3-up genes, 226 DEa E4-up genes, 232 DEg E3-up genes, 268 DEg E4-up genes. (B) The illustration of the related biological processes of the enriched GO terms in Figure 4A, which occur on embryonic day 3 to embryonic day 4 of the human preimplantation embryos. EGA: embryonic genome activation.

DISCUSSION

Differential expression analysis is one of the most important aspects for scRNA-seq data analyses (3,6,25). The mixture of real zeros and dropout zeros in scRNA-seq data posed a great challenge to DE genes detection (2,6,9). Due to the special characteristics of scRNA-seq data, traditional DE analysis methods developed based on bulk RNA-seq data are not suitable for the analysis of single-cell data (19). Existing DE analysis methods designed for scRNA-seq data are also not capable of distinguishing the two types of zeros (19,25,34,35), and may therefore miss important indicators of gene expression status in the cells.

We proposed a method called DEsingle which could estimate the proportions of the two kinds of zeros and then detect the DE genes based on ZINB model. We studied the mRNA capture procedure using a mathematical model. Under the random capture assumption, we proved that after the mRNA

capture procedure, the data changed from a ZINB distribution to another ZINB distribution, with parameter p changing to p^* while all the other parameters remain unchanged. Based on this, the unchanged parameter θ in our model represents the proportion of real zero expression of a gene in a group of cells. This is also intuitively understandable in the sense that real zeros will still be zeros after the capture procedure, and its proportion in the cells will not change due to the capture. On the contrary, the parameter p of NB part in the ZINB model becomes smaller after the random capture, resulting in a smaller mean of the data and a higher probability of generating dropout zeros. That is to say, the dropout zeros produced by mRNA capture procedure are taken into account by the NB part of the ZINB model. Therefore, we could distinguish the real zeros and dropout zeros by estimate the parameter θ, which is the proportion of real zeros in the data.

In the model, the parameter θ represents the ratio of cells in which the gene is not active in its expression, i.e., the proportion of cells in which the expression status is “OFF”. When a gene’s θ has significant difference between two groups of cells, the gene’s expression statuses of the two groups of cells are different. Therefore, the two parts in our model represents two types of differential expression: differential expression status of the gene in the two groups or differential expression abundance between the two groups. And these two cases can also happen at the same time, which we refer to as general differential expression. These three types of differential expression represent differential biological situations. We showed in this study that the DEsingle method we proposed can not only detect differential expression with higher accuracy, but also can differentiate these three types of differential expression, which we named as different expression status (DEs), different expression abundance (DEa) and general differential expression (DEg). We applied the method on a human preimplantation embryonic cell scRNA-seq dataset and successfully detected genes with differential expression status and/or abundances between E3 and E4 cells with enriched GO functions that are specific to the biological processes at E3 or E4.

AVAILABILITY

An R package DEsingle for differential expression analysis of scRNA-seq data is available in the GitHub repository (https://github.com/miaozhun/DEsingle)

SUPPLEMENTARY DATA

Supplementary Data are available at bioRxiv online.

ACKNOWLEDGEMENTS

The authors greatly acknowledge the helpful discussions with Drs. Ke Deng, Xiaowo Wang and Jun Li on this work.

FUNDING

This work was supported in part by the National Natural Science Foundation of China and the Human Cell Atlas pilot project of the Chan Zuckerberg Initiative.

Conflict of interest statement. None declared.

REFERENCES

1. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. and Teichmann, S.A. (2015) The technology and biology of single-cell RNA sequencing. *Mol Cell*, **58**, 610-620.
2. Saliba, A.E., Westermann, A.J., Gorski, S.A. and Vogel, J. (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res*, **42**, 8845-8860.
3. Wang, Y. and Navin, N.E. (2015) Advances and applications of single-cell sequencing technologies. *Mol Cell*, **58**, 598-609.
4. Tang, F., Lao, K. and Surani, M.A. (2011) Development and applications of single-cell transcriptome analysis. *Nature methods*, **8**.
5. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol*, **17**, 13.
6. Stegle, O., Teichmann, S.A. and Marioni, J.C. (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*, **16**, 133-145.
7. Bacher, R. and Kendziorski, C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol*, **17**, 63.
8. Grun, D. and van Oudenaarden, A. (2015) Design and analysis of single-cell sequencing experiments. *Cell*, **163**, 799-810.
9. Poirion, O.B., Zhu, X., Ching, T. and Garmire, L. (2016) Single-cell transcriptomics bioinformatics and computational challenges. *Front Genet*, **7**, 163.
10. Eberwine, J., Sul, J.-Y., Bartfai, T. and Kim, J. (2013) The promise of single-cell sequencing. *Nature Methods*, **11**, 25-27.
11. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol*, **11**, R106.
12. Miao, Z. and Zhang, X. (2016) Differential expression analyses for single-cell RNA-Seq: old questions on new data. *Quantitative Biology*, **4**, 243-260.
13. Jaakkola, M.K., Seyednasrollah, F., Mehmood, A. and Elo, L.L. (2016) Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief Bioinform*.
14. Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D. and Betel, D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*, **14**, R95.
15. Seyednasrollah, F., Laiho, A. and Elo, L.L. (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform*, **16**, 59-70.
16. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.
17. Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136-138.
18. Shapiro, E., Biezuner, T. and Linnarsson, S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet*, **14**, 618-630.
19. Delmans, M. and Hemberg, M. (2016) Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, **17**, 110.
20. Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C. *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods*, **10**, 1093-1095.
21. Macaulay, I.C. and Voet, T. (2014) Single cell genomics: advances and future perspectives. *PLoS Genet*, **10**, e1004126.
22. Boon, W.C., Petkovicduran, K., Zhu, Y., Manasseh, R., Horne, M.K. and Aumann, T.D. (2011) Increasing cDNA yields from single-cell quantities of mRNA in standard laboratory reverse

transcriptase reactions using acoustic microstreaming. *Journal of Visualized Experiments Jove*, e3144.

23. Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M. and Wold, B.J. (2014) From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res*, **24**, 496-510.
24. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lonnerberg, P. and Linnarsson, S. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*, **11**, 163-166.
25. Kharchenko, P.V., Silberstein, L. and Scadden, D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat Methods*, **11**, 740-742.
26. Pierson, E. and Yau, C. (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol*, **16**, 241.
27. Sandberg, R. (2013) Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, **11**, 22-24.
28. Peccoud, J. and Ycart, B. (1995) Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, **48**, 222-234.
29. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. and Tyagi, S. (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, **4**, e309.
30. Garay, A.M., Hashimoto, E.M., Ortega, E.M.M. and Lachos, V.H. (2011) On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, **55**, 1304-1318.
31. Rodrigues, J. (2006) Full bayesian significance test for zero-inflated distributions. *Communications in Statistics: Theory and Methods*, **35**, 299-307.
32. Johnson, W.D., Burton, J.H., Beyl, R.A. and Romer, J.E. (2015) A simple chi-square statistic for testing homogeneity of zero-inflated distributions. *Open J Stat*, **5**, 483-493.
33. Ridout, M., Hinde, J. and Demétrio, C.G. (2001) A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**, 219.
34. Vu, T.N., Wills, Q.F., Kalari, K.R., Niu, N., Wang, L., Rantalainen, M. and Pawitan, Y. (2016) Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, **32**, 2128.
35. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S. and Rinn, J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*, **32**, 381-386.
36. Petropoulos, S., Edsgard, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R. and Lanner, F. (2016) Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell*, **165**, 1012-1026.
37. Hardcastle, T.J. and Kelly, K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.
38. Tse, S.K., Chow, S.C., Lu, Q. and Cosmatos, D. (2009) Testing homogeneity of two zero-inflated Poisson populations. *Biom J*, **51**, 159-170.
39. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1-38.
40. Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, **9**, 60-62.
41. Gillespie, D.T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, **22**, 403-434.
42. Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.B., Lonnerberg, P. and Linnarsson, S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res*, **21**, 1160-1167.

43. Zweig, M.H. and Campbell, G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**, 561-577.
44. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44-57.
45. Niakan, K.K., Han, J., Pedersen, R.A., Simon, C. and Pera, R.A. (2012) Human pre-implantation embryo development. *Development*, **139**, 829-841.
46. Brison, D.R., Sturmey, R.G. and Leese, H.J. (2014) Metabolic heterogeneity during preimplantation development: the missing link? *Hum Reprod Update*, **20**, 632-640.
47. Fleming, T.P., Sheth, B. and Fesenko, I. (2001) Cell adhesion in the preimplantation mammalian embryo and its role in trophectoderm differentiation and blastocyst morphogenesis. *Front Biosci*, **6**, D1000-D1007.
48. Larue, L., Ohsugi, M., Hirchenhain, J. and Kemler, R. (1994) E-cadherin null mutant embryos fail to form a trophectoderm epithelium. *Proceedings of the National Academy of Sciences*, **91**, 8263-8267.
49. Li, C.-B., Hu, L.-L., Wang, Z.-D., Zhong, S.-Q. and Lei, L. (2009) Regulation of compaction initiation in mouse embryo. *Hereditas (Beijing)*, **31**, 1177-1184.
50. Adams, C.L., Chen, Y.-T., Smith, S.J. and Nelson, W.J. (1998) Mechanisms of epithelial cell–cell adhesion and cell compaction revealed by high-resolution tracking of E-cadherin–green fluorescent protein. *The Journal of Cell Biology*, **142**, 1105-1119.
51. Cui, X.S., Li, X.Y., Shen, X.H., Bae, Y.J., Kang, J.J. and Kim, N.H. (2007) Transcription profile in mouse four-cell, morula, and blastocyst: genes implicated in compaction and blastocoel formation. *Molecular Reproduction and Development*, **74**, 133-143.
52. Kim, S.A., Tai, C.-Y., Mok, L.-P., Mosser, E.A. and Schuman, E.M. (2011) Calcium-dependent dynamics of cadherin interactions at cell–cell junctions. *Proceedings of the National Academy of Sciences of U.S.A.*, **108**, 9857-9862.
53. Iwata, K., Yumoto, K., Sugishima, M., Mizoguchi, C., Kai, Y., Iba, Y. and Mio, Y. (2014) Analysis of compaction initiation in human embryos by using time-lapse cinematography. *Journal of Assisted Reproduction and Genetics*, **31**, 421-426.
54. Rasmussen, C.H., Petersen, D.R., Moeller, J.B., Hansson, M. and Dufva, M. (2015) Collagen type I improves the differentiation of human embryonic stem cells towards definitive endoderm. *PLoS One*, **10**, e0145389.
55. Christians, E., Campion, E., Thompson, E.M. and Renard, J.-P. (1995) Expression of the HSP 70.1 gene, a landmark of early zygotic activity in the mouse embryo, is restricted to the first burst of transcription. *Development*, **121**, 113-122.
56. Nelson, D.L. and Cox, M.M. (2005) *Principles of Biochemistry*. New York: WH Freeman and Company.
57. Weaver, R.F. (2008) *Molecular Biology*. New York: McGraw-Hill.
58. Hamatani, T., Carter, M.G., Sharov, A.A. and Ko, M.S. (2004) Dynamics of global gene expression changes during mouse preimplantation development. *Developmental Cell*, **6**, 117-131.
59. Nothias, J.-Y., Majumder, S., Kaneko, K.J. and DePamphilis, M.L. (1995) Regulation of gene expression at the beginning of mammalian development. *Journal of Biological Chemistry*, **270**, 22077-22080.
60. Schultz, R.M. (2002) The molecular foundations of the maternal to zygotic transition in the preimplantation embryo. *Human Reproduction Update*, **8**, 323-331.
61. Dumollard, R., Duchen, M. and Carroll, J. (2007) The role of mitochondrial function in the oocyte and embryo. *Current Topics in Developmental Biology*, **77**, 21-49.
62. Wilding, M., Coppola, G., Dale, B. and Di Matteo, L. (2009) Mitochondria and human preimplantation embryo development. *Reproduction*, **137**, 619-624.