Computational and Structural Biotechnology Journal 20 (2022) 4850–4859

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mathematical model for the relationship between single-cell and bulk gene expression to clarify the interpretation of bulk gene expression data

Daigo Okada *, Cheng Zheng, Jian Hao Cheng

Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, South Research Bldg. No.1(5F), 53 Shogoinkawahara-cho, Sakyo-ku, Kyoto 6068507, Kyoto, Japan

ARTICLE INFO

Article history:
Received 14 July 2022
Received in revised form 26 August 2022
Accepted 26 August 2022
Available online 5 September 2022

Keywords:
Gene expression
Cellular heterogeneity
Single cell
Probability distribution
Differential expression analysis
Differential variability analysis

ABSTRACT

Background. Differential expression analysis is a standard approach in molecular biology. For example, genes whose expression levels differ between diseased and non-diseased samples are considered to be associated with that disease. On the other hand, differential variability analysis focuses on the differences of the variances of gene expression between sample groups. Although differential variability is also known to capture biological information, its interpretation remains unclear and controversial. Recent single-cell analyses have revealed that differences between sample groups can affect gene expression in a cellular subset-specific manner or by altering the proportion of a particular cellular subset. The aim of this study is to clarify the interpretation of mean and variance of bulk gene expression data.

Method. We developed a mathematical model in which the bulk gene expression value is proportional to the mean value of the single-cell gene expression profile. Based on this model, we performed theoretical, simulated and real single-cell RNA-seq data analyses.

Result and Conclusion. We identified how differences in single-cell gene expression profiles affect the differences in the mean and the variance of bulk gene expression. It is shown that differential expression analysis of bulk expression data can overlook significant changes in gene expression at the single-cell level. Further, differential variability analysis capture the complex feature affected by different gene expression shifts for each subset, changes in the proportions of cellular subsets, and variation in single-cell distribution parameters among samples.

1. Introduction

Performing differential expression (DE) analysis of different sample groups is a standard approach in molecular biology. In recent years, transcriptome data have been used to comprehensively identify DE genes in different experimental groups [1], and several bioinformatics methods have been developed for this purpose [2–4]. In DE analysis, if gene expression levels differ significantly between diseased and non-diseased donors, then the genes are considered to be associated with that disease. Similarly, a comparison of gene expression data from different tissues or anatomical regions can therefore be used to identify tissue/region-specific genes [5–7]. Expression quantitative trait locus analysis (eQTL) can be used to identify genetic variants in genotype groups that are significantly associated with gene expression levels, and in so doing, can facilitate an understanding of the mechanisms underlying gene regulation and interpretations of functional genetic variants [8,9]. Specifically, these DE analyses identify differences in mean expression values for bulk gene expression data between the groups based on disease, tissue or genotype (Fig.1(a)).

On the other hand, differential variability (DV) analysis is another approach for identifying differences in gene expression [10]. DV analysis captures differences in variance of gene expression values between the groups (Fig.1(a)). DV can capture biological information about a target disease or trait. To date, studies have employed DV analysis of transcriptome data to provide biological insights about disease and aging [11–14]. For example, a strong relationship has been reported between variability in gene expression and a chronic lymphocytic leukemia subtype [14]. In the context of eQTL analysis, the genetic loci associated with variance in bulk gene expression value are discussed as expression variability QTL (evQTL) [15]. Although the biological processes underlying DV have been investigated from a biological standpoint, such as gene expression noise or epigenetics [16], the interpretation of DV of gene expression remains unclear and controversial.

* Corresponding author.
E-mail address: dokada@genome.med.kyoto-u.ac.jp (D. Okada).

https://doi.org/10.1016/j.csbj.2022.08.062

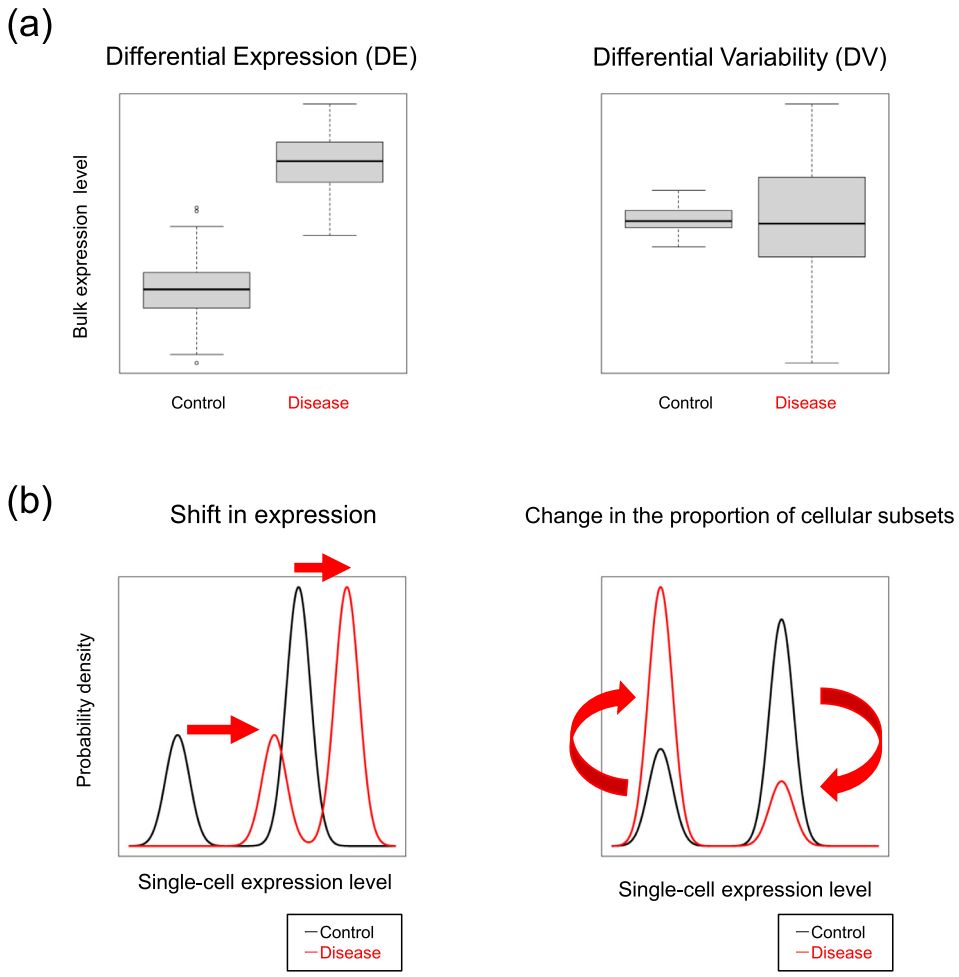

Fig. 1. Background of the study. (a) Concept of differential expression (DE) and differential variability (DV) in gene expression analysis. DE and DV analysis capture the mean and the variance of bulk expression distributions between groups, such as control vs. disease groups, respectively. (b) Graphical illustrations showing changes in the distribution of two types of single-cell expression; the first type is characterized by a shift in expression levels in each cellular subset (left). The magnitude and direction of the shift can differ for each cellular subset, and can be expressed as a change in the individual distributions that make up the single-cell expression distribution. The second type is a change in the cellular subset proportion (right). This can be expressed as a change in the proportion of the component distribution.

Due to cellular heterogeneity, bulk gene expression data in DE or DV analyses are not typically sufficient for capturing the changes in the gene expression profiles of a cell population. Each sample in an experiment contains randomly selected cells, and each cell has a different gene expression level. Consequently, the cell population profile can be expressed as a probability distribution of gene expression, and the bulk gene expression data captures information of the average value of this distribution [17]. Recent advances in single-cell analysis have reported the existence of two different types of changes in a single-cell expression profile; shifts in gene expression and changes in the proportion of cellular subsets.

First, group differences shift the level of gene expression in the cell, and the direction and magnitude of this shift depends on the cellular subset. For example, it is known that tumor cell subpopulations show distinct drug responses [18]. The recent studies combining single nucleotide polymorphism (SNP) genotype data with single-cell RNA-seq (scRNA-seq) data or cytometry has shown that the effects of genetic variants on gene expression may differ depending on cellular subsets [19–21]. Such changes in the single-cell expression distribution can be expressed as shown in the left panel of Fig.1(b).

Second, group differences change the proportion of cellular subsets, as shown in the right panel of Fig.1(b). Differences among groups can alter proportion of cellular subsets by affecting cell differentiation, maturation and transformation. Studies have been conducted to identify cellular subsets with different proportions between sample groups [22,23]. Previous studies combining SNP genotype and cytometry analyses identified SNPs associated with different lymphocyte subsets [24,25]. In addition, it has been suggested that a large number of SNPs are associated with individual differences in lymphocyte profiles, even though their effects are small [25]. Changes in the proportion of cellular subsets can affect the bulk gene expression value. For example, when the proportion of a cellular subset with a relatively high gene expression level increases, the bulk expression levels also increase.

In the complex physiological and pathological changes that occur within cells, both shifts in gene expression and changes in cellular subset proportions can occur simultaneously, resulting in a combined contribution to the bulk expression value. For example, both of them in intestinal epithelial cell population have been observed in patients with Crohn's disease [26]. Evaluating how these changes in single-cell expression profiles are manifested in DE and DV genes is central to understanding the biological mechanisms underlying both DE and DV analysis. Further, given the increased interest in single-cell expression analysis in recent years, it is important to evaluate the results of single-cell expression analyses and compare them to the results of bulk expression analyses that have been reported to date.

In this study, we describe a mathematical model for examining bulk gene expression levels and the single-cell expression distribution behind them. Specifically, single-cell expression profiles are modeled using a mixed probability distribution, and the relationships among their parameters and the mean and variance of the bulk expression values among samples are clarified. The model proposed in this study clarifies the interpretation of DE and DV analysis and provides new insights into the relationship between bulk and single-cell data analyses.

2. Methods

2.1. Mathematical model for relationship between a single-cell expression profile and a bulk gene expression value

We developed the following mathematical model to clarify the relationship between one single-cell expression profile and its bulk gene expression value for one gene (Fig.2(a)). The bulk samples of multicellular organisms consist of many cells, which show cellular

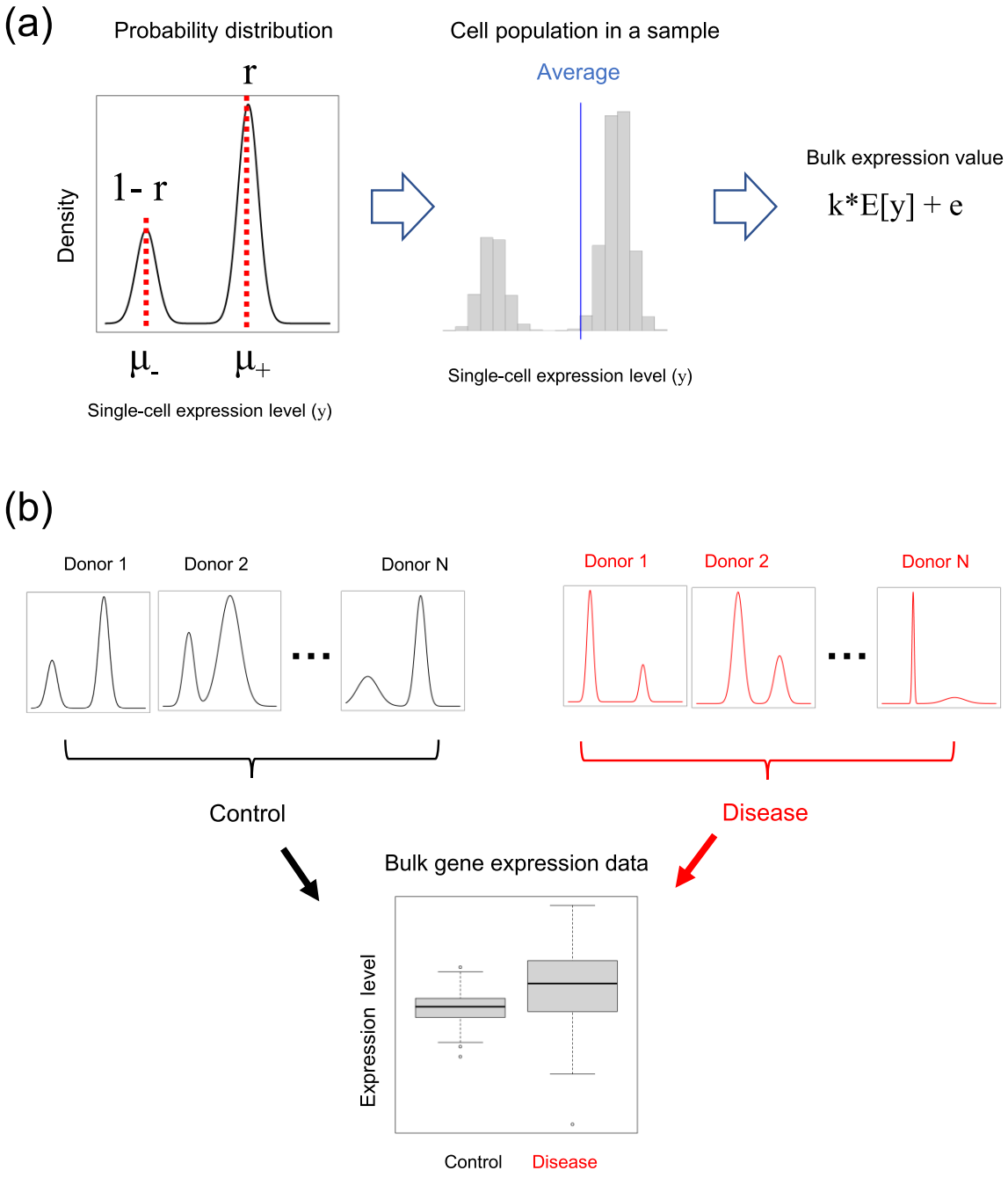

Fig. 2. Mathematical model of the relationship between single-cell expression profiles and bulk expression data. (a) Model of the relationship between the single-cell expression profile and bulk gene expression values. The single-cell expression profile is modeled as a mixture distribution. μ_+ and μ_- are the expected values of the component distributions and r is the proportion of cellular subset. Bulk samples are obtained as statistical samples where the number of cells contained is sufficiently large. Bulk gene expression values are proportional to the expected value of the population distribution + the measurement error ($k * E[y] + e$). (b) Model of the bulk expression data analysis. In the bulk data analysis, multiple sample values are taken where μ_+, μ_-, and r have different values in each sample. This mathematical model evaluates the relationships among the statistics of μ_+, μ_- and r, and the bulk gene expression Y.

heterogeneity and consist of multiple cellular subsets. The distribution of gene expression for these cells can be expressed as a mixture distribution of those of different cellular subsets.

To evaluate the effect of a specific cellular subset, we used cellular subset + and cellular subset-. The single-cell expression value for a cell of cellular subset + is assumed to follow a probability distribution $f_+(y)$ with mean μ_+. Similarly, that for a cell in cellular subset- is assumed to follow a probability distribution $f_-(y)$ with mean μ_-. Let the proportion of cellular subset + be r and that of cellular subset- be $1-r$. Due to genetic and environmental effects, the parameters affecting the single-cell expression distributions are assumed to vary among donors. We modeled μ_+, μ_-, r by assuming that these are random variables that are independent of each other.

Under this model, the single-cell expression level y follows a mixture probability distribution, as follows:

$$y \sim rf_+(y) + (1-r)f_-(y) \tag{1}$$

The bulk gene expression value (Y) can be considered as the mean value for a single-cell expression distribution. If the bulk sample contains a sufficient number of cells, then the bulk gene expression level (Y) can be modeled with a proportionality constant k and measurement error (e) as follows:

$$\begin{aligned} Y &= kE[y] + e \\ &= kr\mu_+ + k(1-r)\mu_- + e \end{aligned} \tag{2}$$

where, e is assumed to be independent of other random variables μ_+, μ_-, r.

For simplicity, we set k = 1 in the discussion.

$$Y = r\mu_+ + (1-r)\mu_- + e$$

As a result, the group difference in bulk expression data, such as control vs. disease, are interpreted via the single-cell gene expression distribution according to Eq. 3. In DE analysis, the bulk data for the disease group (Group D) and the control group (Group C), $E[Y^D] - E[Y^C]$ can be detected statistically as the difference of bulk expression values for each group. This model allows us to mathematically evaluate the relationship between the parameters for the single-cell expression distribution with cellular heterogeneity and the results of the bulk gene expression analysis (Fig.2(b)).

2.2. Relationships among cellular heterogeneity, DE and DV analyses in the bulk experiment

From Eq. 3, the bulk expression value Y can be written as

$$E[Y] = E[r]E[\mu_+] + (1-E[r])E[\mu_-] \tag{4}$$

Consider the case of comparing bulk data for disease group (Group D) and the control group (Group C). In Group D, the expected value for the parameters of single-cell expression distribution in the model is shifted from those of group C, as follows:

$$\begin{aligned} E[\mu_+^D] &= E[\mu_+^C] + \Delta\mu_+ \\ E[\mu_-^D] &= E[\mu_-^C] + \Delta\mu_- \\ E[r^D] &= \alpha_r E[r^C] \end{aligned} \tag{5}$$

where $\Delta\mu_+, \Delta\mu_-$ and α_r express the difference in μ_+, μ_- and r between groups. Note that $\alpha_r E[r^C]$ is restricted to the range 0 to 1. The differential expression between the two groups, i.e., $E[Y^D]$ - $E[Y^C]$, can be expressed as follows, based on Eq. 4 (see Appendix A for details).

$$E[Y^D] - E[Y^C] = (E_+^C - E_-^C)(E_r^C(\alpha_r - 1) + \alpha_r E_r^C d_+ + (1-\alpha_r E_r^C)d_-)$$

where, $d_+ = \frac{\Delta\mu_+}{(E_+ - E_-)}$ and $d_- = \frac{\Delta\mu_-}{(E_+ - E_-)}$. In addition, we let E_+^C, E_-^C and E_r^C equivalent to $E[\mu_+], E[\mu_-]$ and $E[r]$ in Group C, respectively.

Depending on the combination of α_r, d_+, and $d_-, E[Y^D] - E[Y^C]$ can take a value of zero and never be identified by DE analysis, even though positive activation of gene expression is occurring at the single-cell level. Based on these results, if a group difference affects both the cellular subset proportion and the gene expression level in each cell, then it can be missed in the bulk gene expression analysis.

Indeed, if $d_+ > 0$ and $d_- > 0$, then the condition for $E[Y^D] - E[Y^C] \le 0$ can be expressed as follows (from Eq. 6):

$$\alpha_r \le \frac{1 - \frac{d_-}{E_r^C}}{1 + d_+ - d_-} \tag{7}$$

where we assumed $E_+^C > E_-^C, r > 0, 1 + d_+ - d_- > 0$. $E[Y^D] = E[Y^C]$ is satisfied when the equal sign holds.

On the other hand, variance in the bulk gene expression value can be calculated from Eq. 3, as follows:

$$V[Y] = V[r\mu_+ + (1-r)\mu_- + e] \tag{8}$$

From Eq. 8, the following relationship can be mathematically derived (see Appendix B for details):

$$\begin{aligned} V[Y] = (E_+ - E_-)^2 V_r + E_r^2 V_+ + (1-E_r)^2 V_- + (V_+ + V_-)V_r \\ + V[e] \end{aligned} \tag{9}$$

In addition, we let E_+, E_-, E_r, V_+, V_- and V_r be equivalent to $E[\mu_+], E[\mu_-], E[r], V[\mu_+], V[\mu_-]$ and $V[r]$, respectively. Eq. 9 suggests that there are three main factors that explain the differential variability between groups. First, V_+, V_- and V_r directly affect the variability in bulk gene expression. Second, the change in E_r affects the variability in bulk gene expression via the term $E_r^2 V_+ + (1-E_r)^2 V_-$. If the group difference increases the more variable subset proportion, then $V[Y]$ can be increased. Third, $(E_+ - E_-)^2$ can affect $V[Y]$. If the group difference changes $(E_+ - E_-)^2$, then it affects the bulk expression variance via the term $(E_+ - E_-)^2 V_r$. Importantly, even if V_-, V_+ and V_r do not change, a change in E_+, E_- and E_r can change the variance in the bulk gene expression. Note that cell-to-cell variability $V[y]$ does not appear in this equation.

2.3. Visualization of the difference between DE and DV genes

Based on the above theory, we visualized the relationship between the parameter $(d+, d-, \alpha_r)$ and an increase or decrease in the mean and variance of the bulk gene expression. We focused on a situation where expression is increased in both cellular subsets ($d_+ \ge 0, d_- \ge 0$) and the proportion of cellular subset + is decreased ($\alpha_r \le 1$). For the combinations of equally spaced d + and d- (0, 0.1, 0.2, …, 1) and three patterns of α_r (0.2, 0.5, 0.8), we calculated the difference in the means and variances, and visualized the parameter space with an increase and a decrease in the mean and the variance of the bulk gene expression. We set other parameters in the model as follows: $E_+ = 2, E_- = 1, E_r = 0.5, V_+ = 0.3, V_- = 0.1$ and $V_r = 0.05$.

2.4. Simulation analysis of DV genes

We created a computational simulation scheme for the proposed mathematical model. First, this model contains fourteen parameters: N is the number of samples, $E_+, V_+, E_-, V_-, E_r, V_r, d_+, d_-$ and α_r are the parameters that define individual differences in the distribution parameters described in

the above model, n is the number of cells in the bulk sample, V_{err} is the measurement noise associated with quantifying the bulk expression value, and V_{y+} and V_{y-} are the variances of single-cell expression levels in each subset.

First, for N samples in the Group C, μ_+, μ_- and r are sampled from the Normal or uniform distributions as shown below, and assigned to each sample. Normal and uniform distributions are uniquely determined by the mean and variance parameters.

$$\mu_+ \sim Normal(E_+, V_+)$$
$$\mu_- \sim Normal(E_-, V_-)$$
$$r \sim Uniform(E_r, V_r)$$

Also, for N samples in the Group D, μ_+, μ_-, r are sampled according to d_+, d_- and α_r, based on Eq. 5.

$$\mu_+ \sim Normal(E_+ + d_+, V_+)$$
$$\mu_- \sim Normal(E_- + d_-, V_-)$$
$$r \sim Uniform(\alpha_r E_r, V_r)$$

Next, we generated the single-cell expression value (y) of each sample based on Eq. 1, where $f_+(y)$ and $f_-(y)$ are modeled as normal distributions. From the following formula, we sampled n cells independently.

$$y \sim rNormal(\mu_+, V_{y+}) + (1-r)Normal(\mu_-, V_{y-})$$

After sampling n single-cell expression values, the bulk expression value is calculated as the mean of $\boldsymbol{y} = [y_1, y_2 \ldots y_n]$ and a measurement error (e) is added.

$$Y = \bar{\boldsymbol{y}} + e$$
$$e \sim Normal(0, V_{err})$$

As a result, the bulk expression values for N control samples and N disease samples were simulated.

Based on this simulation scheme, we simulated two situations in which only the variance of the bulk expression is changed without changing the mean value. Eq. 9 shows that the difference in E_r (Example 1) or $(E_+ - E_-)^2$ (Example 2) can change the bulk expression variance, even if V_+, V_- and V_r are the same in the two groups.

(Example 1) We simulated an example where the variance in bulk gene expression decreases as E_r decreases in the disease group when $V_+ > V_-$. The model parameters were as follows: $N = 1000, E_+ = 2, E_- = 1, E_r = 0.9, V_+ = 1, V_- = 0.1, V_r = 0.001$, $n = 10000, V_{err} = 0.1, V_{y+} = 1$ and $V_{y-} = 1$. $d_+, d_- = 0.5$ were used so that the value of $E_+ - E_-$ was the same in the control and disease groups. α_r was set so that $E[Y^C] = E[Y^D]$was satisfied. We checked the simulated bulk expression data for the two groups to confirm whether the changes in the proportions of cellular subsets could cause DV.

(Example 2) We simulated an example where the variance in the bulk gene expression increases as $E_+ - E_-$ increases. The model parameters were as follows: $N = 1000, E_+ = 2, E_- = 1, E_r = 0.9, V_+ = 0.01, V_- = 0.01, V_r = 0.001, n = 10000$, $V_{err} = 0.1, V_{y+} = 1$ and $V_{y-} = 1$. V_+ and V_- were the same to repress the effect of differences in the proportions of different cells. Instead, $d_+ = 10$ and $d_- = 0.1$ were used so that $E_+ - E_-$ increases. α_r was determined so that $E[Y^C] = E[Y^D]$was satisfied. We checked the simulated bulk expression data for the two groups to confirm whether the different gene expression shifts for each subset could could cause DV.

2.5. Real single-cell RNA-seq data analysis

Here, we described the analysis of real single-cell RNA-seq data by applying our method to elucidate single-cell expression changes underlying bulk DE and DV genes. We used the public scRNA-seq dataset for ulcerative colitis (UC) (NCBI GEO ID:GSE125527) [27]. In this study, we used data for the processed scRNA-seq of human peripheral blood mononuclear cells (PBMCs) from seven patients with ulcerative colitis (UC) and eight control donors (NCBI GEO ID:GSM3576411-GSM3576425). After log(1 + count value) transformation, the sum of the gene expression values for each cell was normalized to be 10^6 and used for downstream analysis.

The DE genes and DV genes were identified at the bulk level using the following procedure. Bulk gene expression levels for each sample were defined as the average of the single-cell expression values among cells. For the genes with an average bulk gene expression level of > 20 in the 15 samples, we performed Welch's t-test and F-test analyses to compare the UC and the control groups and calculated p values. We applied Benjamini-Hochberg (BH) correction [28] to the t-test p values and identified the genes with adjusted p values < 0.05 as DE genes. We also applied BH correction to the F-test p values and identified the genes with adjusted p values < 0.05 as DV genes.

For the identified top DE gene and top DV gene with the smallest p value, we estimated the parameters of single-cell expression distribution (μ_+, μ_- and r). In many cases, the single-cell expression distribution consists of the cell population with zero or small expression values, and cell populations with higher expression values. For the single-cell expression distribution of each gene in each sample, we applied kmeans clustering and divided the cells into subset + and subset-. By calculating the mean expressions and proportions of subset + and subset- cells, we could then estimate μ_+, μ_- and r for each sample. We then evaluated the distributions of these estimated parameters for the samples and identified differences in the single-cell expression distributions that underlie the bulk expression of DE and DV genes.

3. Results

3.1. Theoretical and simulation analyses

Fig. 3(a) shows $E[Y^D] - E[Y^C]$ for the combination of d_+, d_- and α_r. Even though the expression shift is positive for subsets ($d_+ \geq 0, d_- \geq 0$), the difference in mean expression value can be negative depending on α_r. Fig.3(b) shows the difference in variance for the combination of d_+, d_-, and α_r. Compared to Fig.3(a), the plot pattern is different. These results indicate that DE and DV analysis capture different features of the single-cell distribution.

We simulated situations where the variance in bulk expression differed between sample groups without changing mean values. Fig.4(a) shows an example where only changes in the proportions of cellular subsets alters variance of bulk gene expression. This is because the subset + cells with large variance decrease in number and the subset- cells with small variance increase in number. Fig. 4, (b,c) shows the parameter distributions for μ_+, μ_- and r in the two groups in this simulation. Fig.4(d-f) shows another example where a shift in gene expression changes variance of bulk gene expression. These results show that a changes in the proportions of cellular subsets or a different expression shifts for cellular subsets can cause DV.

3.2. Real single-cell RNA-seq data analysis

We used a public human scRNA-seq dataset compiled using data for seven patients with UC and eight control subjects. The results of a bulk level analysis showed that there were 289 DE genes and four DV genes. No matches were observed between these DE and DV genes. We investigated the top DE and DV genes with smallest p value in detail.

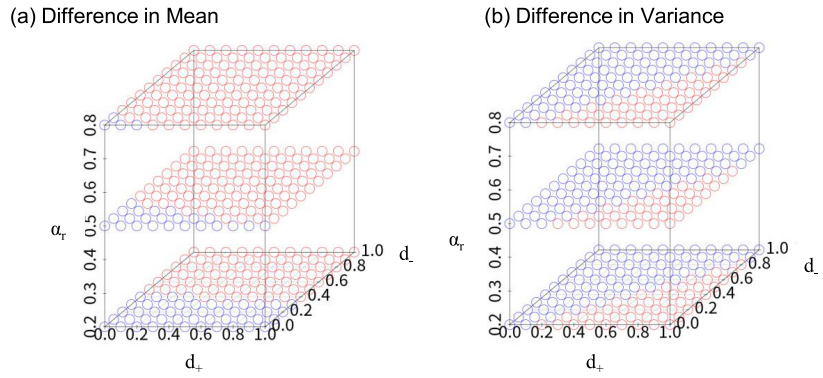

Fig. 3. Example of theoretical analysis involving the comparison of the bulk mean and variance between two groups. (a) Example of plotting difference in mean expression for each combination of d_+, d_- and α_r. Combinations where $E[Y^D] > E[Y^C]$ are plotted in red and combinations where $E[Y^D] \leq E[Y^C]$ are plotted in blue. Even if $d_+ > 0, d_- > 0$, it is possible that $E[Y^D] \leq E[Y^C]$. (b) Example of plotting difference in variance for each combination of d_+, d_- and α_r. Combinations where $V[Y^D] > V[Y^C]$ are plotted in red and combinations where $V[Y^D] \leq V[Y^C]$ are plotted in blue. DE and DV analysis capture different features of the single-cell distribution.

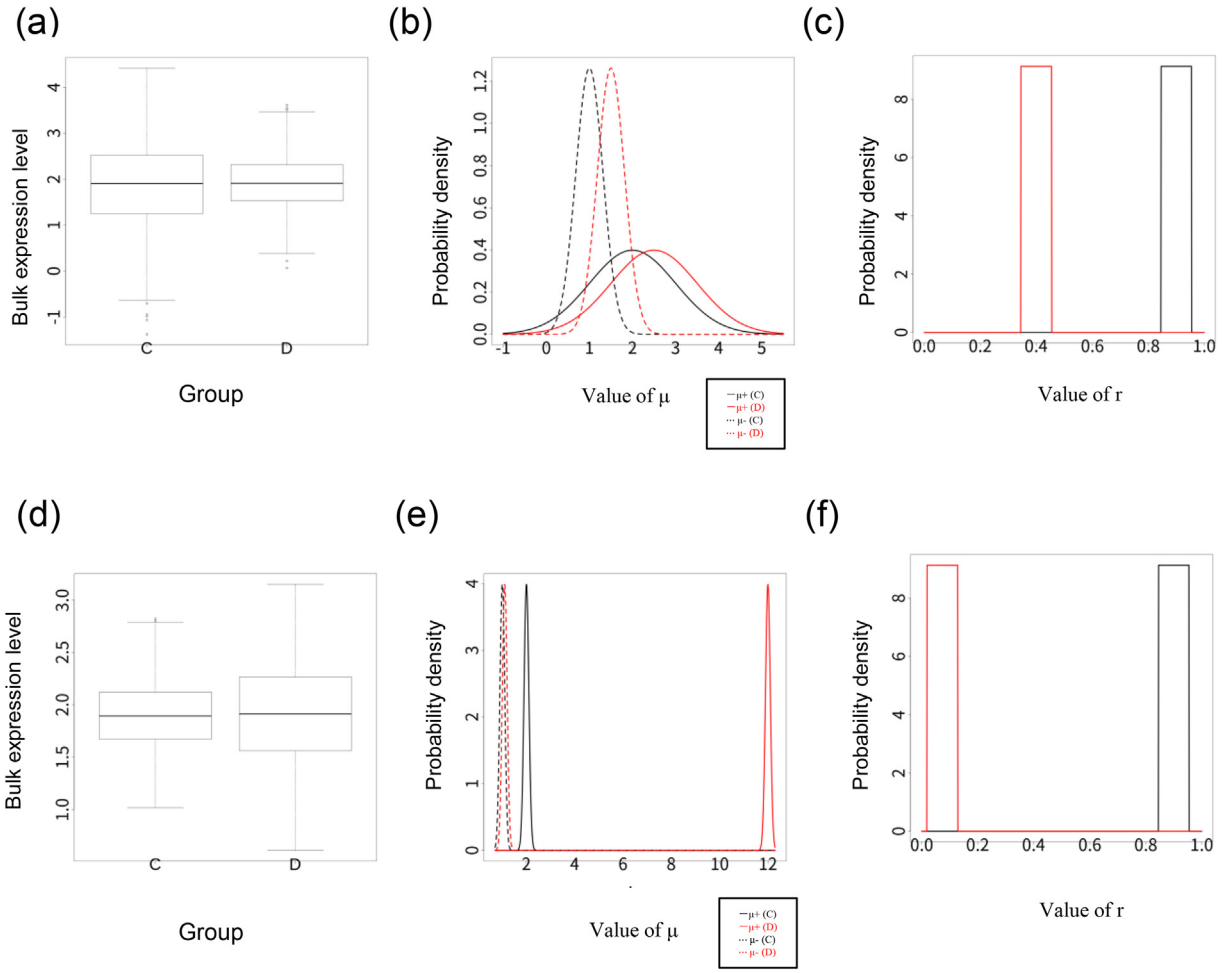

Fig. 4. Results of the simulation analysis. (a) Bulk expression result with no mean difference and a significant difference in variance in Example 1 of simulation analysis. (b) and (c) are the distributions of μ_+, μ_- and r among samples in this simulation. (d-f) are the those of Example 2.

The top DE gene in the bulk expression analysis was *POM121*, which had a mean bulk expression value that decreased significantly in UC group (Fig.5(a); t-test p value is $1.51 * 10^{-5}$). Fig.5(b) shows a histogram of single-cell expression values for pooled cells of each group, which shows that the expression level for this gene was low in a large number of cells (cellular subset -) and high in a small number of cells (cellular subset+). Fig.5, (c,d,e) shows the estimated μ_+ and μ_-, r values for each sample group. The findings suggest that the gene expression level increased markedly in subset + cells in UC group, while the proportion of these cells decreased. Although these two effects act in opposite directions, the greater influence of the former effect increases bulk gene expression.

The top DV gene is *MAP1LC3B2* whose bulk expression variance is significantly decreased in UC group (Fig.6(a); the F-test p value is

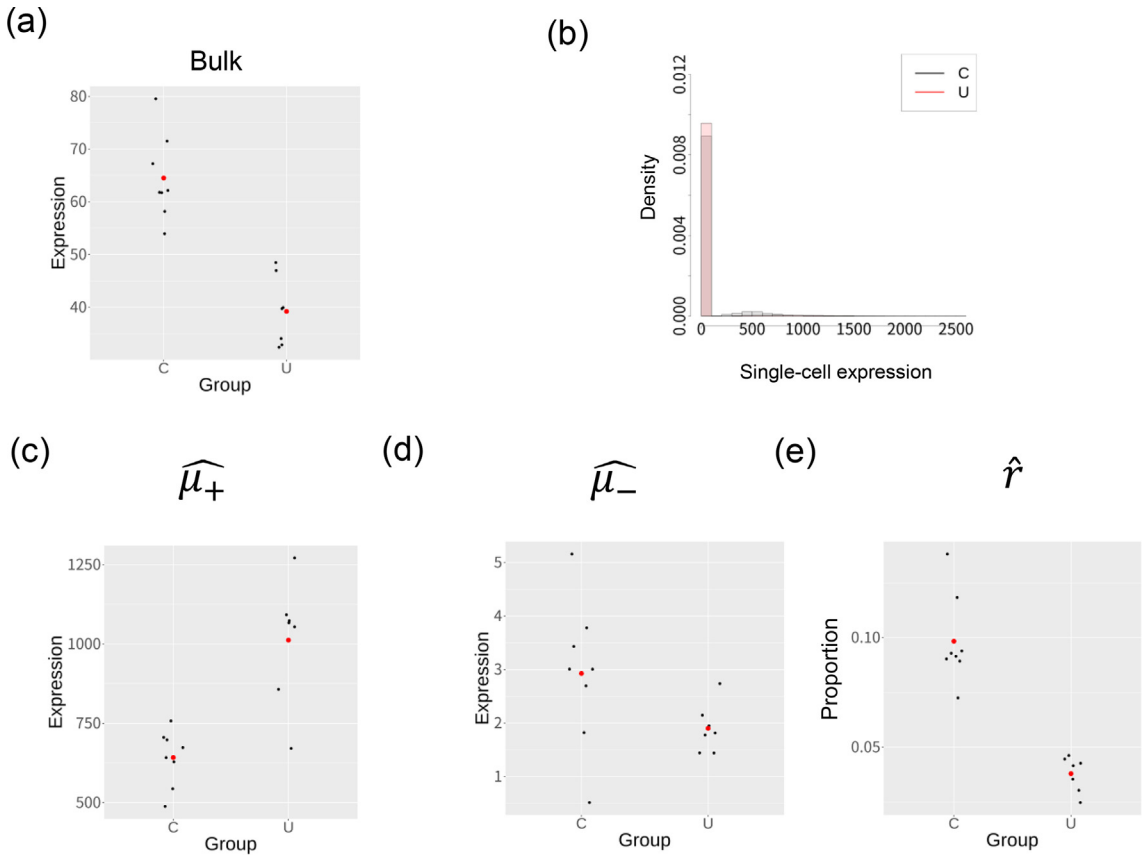

Fig. 5. Analysis of top DE gene with the smallest p value (*POM121*). (a) Bulk gene expression levels of control (C) and ulcerative colitis (U) groups. The red points in the plot represent the mean of the bulk expression values for each group. (b) A histogram showing single-cell expression levels for the C and U groups, plotted for the pooled cells in all samples in each group. (c, d, e) Estimated parameters for single-cell expression distributions ($\hat{\mu_+}$, $\hat{\mu_-}$ and $\hat{r}$) for the samples. Red points represent the mean values obtained for each group.

$8.77 * 10^{-7}$). Fig.6(b) shows a histogram of single-cell expression values for pooled cells of each group, which also shows that the expression level for this gene was low in a large number of cells (cellular subset -) and high in a small number of cells (cellular subset+). Fig.6, (c,d,e) shows the estimated μ_+ and μ_-, r values for each sample group. While the mean and variance of the bulk expression level decreased in UC group, μ_+ increased. The large decrease in the mean and variance of r in UC group is considered to lead significant difference of variance.

4. Discussion

The proposed model provides the insights for the interpretations of DE and DV genes identified using bulk data. First, when changes in the proportions and expression shifts in cellular subsets occur, they may cancel each other out and not be detected in bulk data analysis. Second, underlying DV gene, there is a combined contribution of different expression shifts for cellular subsets, changes in the proportion of cellular subsets, and changes in individual differences in the parameters for single-cell expression profiles. These are important considerations when interpreting the results obtained from bulk expression analysis. Third, DV and DE capture different aspects of single-cell expression profile differences. In recent years, methods for directly detecting dissimilarities in single-cell expression distributions using scRNA-seq or cytometry data have been proposed [29–33]. Our findings provide an insight into the theoretical relationships among DE and DV gene identified in bulk experiment, and differential distributed genes identified in single-cell experiment.

Here we examine the difference between variability in the bulk expression data ($V[Y]$) and the cell-to cell variability ($V[y]$). Recently, studies on single-cell analysis have examined cell-to-cell variability and the statistical methods for analyzing this expression characteristics [34,35]. Our analysis showed that the formula for $V[Y]$ does not include $V[y]$ and that there is no direct relationship between them when the sample contains a sufficiently large number of cells. Nevertheless, shifts in gene expression or changes in the proportion of cellular subsets will also induce changes in $V[y]$ in the same direction in individual samples. In addition, changes in $V[y], V[\mu_+], V[\mu_-]$ and $V[r]$ due to disease or aging may indicate the existence of a common mechanism for disruption of the control mechanism for biological phenomena.

The scRNA-seq expression data is characterized by the inclusion of many zero values [36]. Then, statistical models that can handle zero-inflation are often used in single cell data analysis. Specifically, it is often modeled by probability distributions such as negative binomial, poisson, zero-inflated negative binomial or zero-inflated poisson [36]. Our model is applicable to any probability distribution including the zero-inflation model. Since the relationship between expectation value of distribution and its parameters is known mathematically for these theoretical distributions, it is possible to calculate the values such as E_+ or V_+ from the mean and variance of the parameter values among samples, which provide the insights about the mean and variance of the bulk gene expression values. If future large scale single cell genomics studies will provide insight into the distribution of the parameters among samples, it will enhance our understanding of the relationship between single cell and bulk data analysis even more.

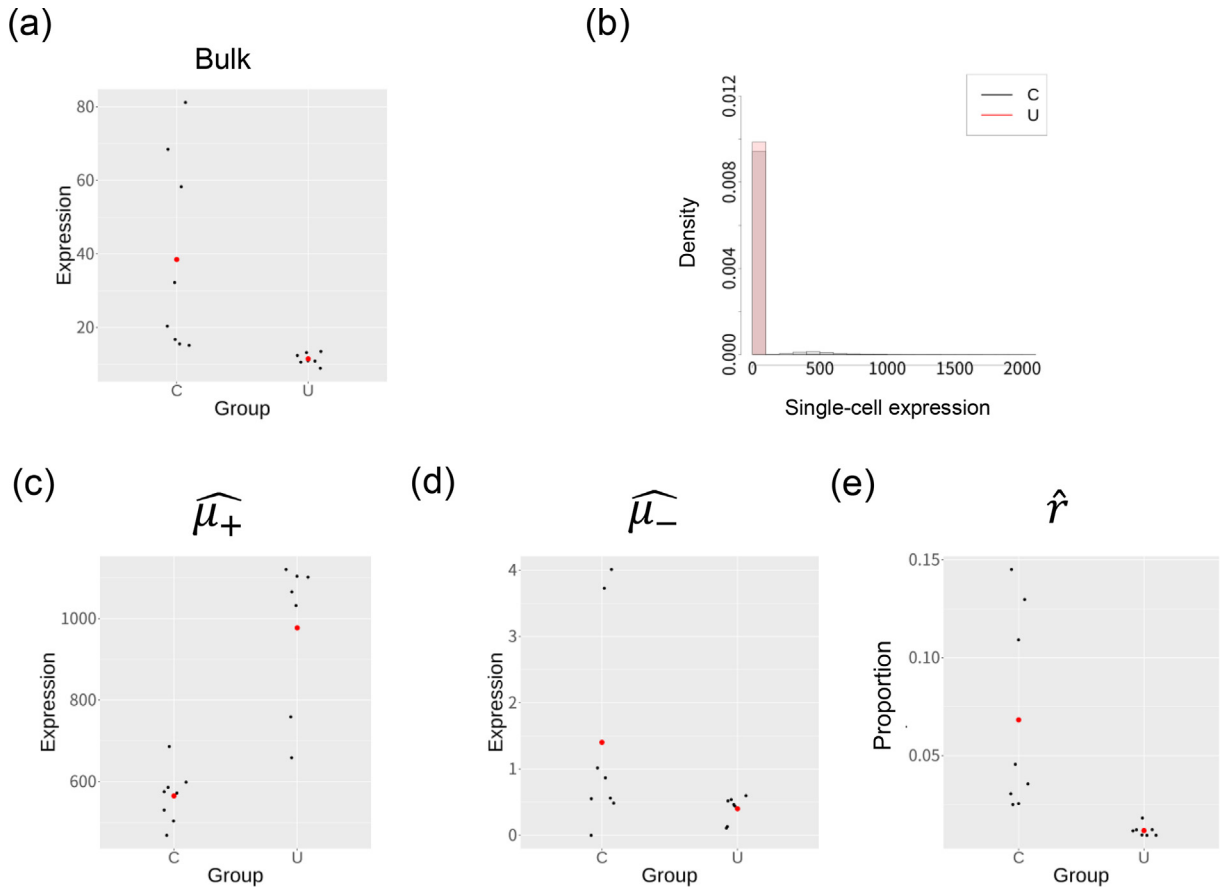

Fig. 6. Analysis of the top DV gene with the smallest p value (*MAP1LC3B2*). (a) The bulk gene expression level of control (C) and ulcerative colitis (U) groups. The red points in the plot represent the mean of the bulk expression values for each group. (b) A histogram showing single-cell expression levels for C and U groups, plotted for the pooled cells in all samples for each group. (c, d, e) Estimated parameters of single-cell expression distributions ($\hat{\mu_+}, \hat{\mu_-}$ and $\hat{r}$) for the samples. Red points represent the mean values obtained for each group.

The analysis of real scRNA-seq data presented in this study is an effective tool for examining the relationship between scRNA-seq data and bulk data, but the study has several limitations. A major limitation is that the estimates obtained for $\hat{E}_+, \hat{E}_-, \hat{E}_r, \hat{V}_+, \hat{V}_-$ and $\hat{V}_r$ are not always accurate. These estimates can be affected by biases associated with clustering and parameter estimation. Also, the number of cellular subsets that are actually present in a tissue is not always known, and the identification and classification of cellular subsets using bioinformatic methods is a major research task in the field of single-cell genomics. Combining the mathematical model proposed in this study with advanced bioinformatics methods may further the study of single-cell genomics.

In our theoretical framework, the assumption that the bulk gene expression value is proportional to the mean value of the single-cell gene expression profile is essential. In real data analysis, the attention should be paid to whether this assumption holds. If very rare subset has non-zero expression value, the analysis will be susceptible to sampling bias. In such cases, it would be necessary to obtain data from a larger number of cells to capture the information of distribution. In addition, this assumption is also an unstable for low expressed gene because gene expression quantification by RNA-seq is unstable technically. Therefore, filtering the low expressed genes are important step as preprocessing under our theoretical framework.

The model described in this study could potentially be used in theoretical fields. Extending the model to poly-genes will allow more bulk expression analysis methods to be applied at the single-cell level. For example, gene co-expression network analysis is performed extensively in transcriptome analysis where it is used to infer biological processes and the roles of important transcription factor genes in complex traits [37,38]. It is necessary to consider at least two genes when using the model to investigate correlations between gene expression. While gene co-expression analysis has been used to clarify relationships in gene regulation, it is unclear what exactly the identified relationships captures. Another extension would be a model that considers spatial information. When acquiring bulk gene expression data, not only information on the shape of the distribution but also spatial information is lost. In recent years, with the development of spatial genomics technology, single cell transcriptome data can be obtained with spatial information [39]. Since our framework works as a model of cellular population heterogeneity in general, it is possible to interpret cellular subset as spatial information. For example, it can also be used as a model for spatial information by setting $f_+(y)$ as the distribution on the specific region and $f_-(y)$ as those on another region. As candidate for future improvements, mathematical models that simultaneously consider cellular subsets and regional information can be considered.

On the application side, our mathematical model could be applied, not only to the analysis of gene expression data, but also to the analysis of arbitrary biomolecular expression data. For example, in the epigenome layer, an increase in the variability of DNA methylation intensity at the bulk level has been reported to be associated with aging [40–43]. In recent years, single-cell

expression data have been obtained for various omics layers. It is considered that the concepts presented in this study will be useful for reinterpreting molecular biology knowledge obtained at the bulk level using single-cell data.

5. Conclusion

In this study, we present a mathematical model to clarify single-cell expression profiles with cellular heterogeneity and bulk gene expression data. The model considered the shift in gene expression and changes in the proportion of cellular subsets. Theoretical and simulation analyses showed that the DE analysis can overlook significant changes in gene expression at the single-cell level. In addition, it is revealed that DV analysis capture the feature affected by different expression shifts for cellular subsets, changes in the proportions of cells, and variations in single-cell distribution parameters among samples. The model presented in this study effectively clarifies the differences in interpretation of DE gene and DV gene identified in bulk experiment and provides new insights into the relationship between bulk data analysis and single-cell data analysis.

6. Funding

This work was funded by a KAKENHI Grant-in-Aid from the Japan Society for the Promotion of Science (JSPS; Grant No. 21K21316).

Code availability

The code used in this study is available athttps://github.com/DaigoOkada/ScBulkModel.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank FORTE Inc. (https://www.forte-science.co.jp/) for proofreading the English in this manuscript.

Appendix A. The detail of Eq. (6)

We calculated $E[Y^D]-E[Y^C]$ using the parameters for the single-cell expression distribution μ_+, μ_- and r. We use the following properties for the expected value of the sum and product of two random variables, A and B.

$$E[A+B]=E[A]+E[B] \tag{A.1}$$

Further, if A and B are independent,

$$E[AB]=E[A]E[B] \tag{A.2}$$

With the above formula, $E[Y^D]-E[Y^C]$ can be expressed as follows. Here, we let E^C_+, E^C_- and E^C_r be equivalent to $E[\mu_+], E[\mu_-]$ and $E[r]$ in control group, respectively.

$$\begin{aligned} & E[Y^D]-E[Y^C] \\ = \;& \alpha_r E^C_r(E^C_+ + \Delta\mu_+) + (1-\alpha_r E^C_r)(E^C_- + \Delta\mu_-) - E^C_r E^C_+ - (1-E^C_r)E^C_- \\ = \;& E^C_r(\alpha_r - 1)(E^C_+ - E^C_-) + E^C_r \alpha_r \Delta\mu_+ + (1-E^C_r\alpha_r)\Delta\mu_- \\ = \;& (E^C_+ - E^C_-)\{E^C_r(\alpha_r-1) + \alpha_r E^C_r \tfrac{\Delta\mu_+}{E^C_+ - E^C_-} + (1-\alpha_r E^C_r)\tfrac{\Delta\mu_-}{E^C_+ - E^C_-}\} \end{aligned}$$

d_+ and d_- can be calculated as follows:

$$d_+ = \tfrac{\Delta\mu_+}{(E^C_+ - E^C_-)}$$
$$d_- = \tfrac{\Delta\mu_-}{(E^C_+ - E^C_-)}$$

where, d_+ and d_- represent the normalized shift in the gene expression value corrected using the difference in the expected value of the gene expression value between the subsets.

As a result, the following Eq. 6 can be derived:

$$E[Y^D]-E[Y^C] = (E^C_+ - E^C_-)(E^C_r(\alpha_r - 1) + \alpha_r E^C_r d_+ + (1-\alpha_r E^C_r)d_-)$$

Appendix B. The detail of Eq. 8

The following properties of the variance and covariance properties of the two random variables A and B are used:

$$V[A+B] = V[A] + V[B] + 2Cov[A,B] \tag{B.1}$$

The covariance can be expressed using expected values, as follows:

$$Cov[A,B] = E[AB] - E[A]E[B] \tag{B.2}$$

The following equation holds for the variance of the product of independent random variables.

$$V[AB] = V[A]V[B] + E[A]^2V[B] + E[B]^2V[A] \tag{B.3}$$

Using these formulas, the variance $V[Y]$ of the bulk expression value can be expressed mathematically as follows, where we let E_+, E_-, E_r, V_+, V_- and V_r be equivalent to $E[\mu_+], E[\mu_-], E[r], V[\mu_+], V[\mu_-], V[r]$ respectively.

$$\begin{aligned} V[Y] = \;& V[r\mu_+ + (1-r)\mu_- + e] \\ = \;& V[r\mu_+ + (1-r)\mu_-] + V[e] \\ = \;& V[r\mu_+] + V[(1-r)\mu_-] + 2Cov[r\mu_+, (1-r)\mu_-] + V[e] \end{aligned} \tag{B.4}$$

$V[r\mu_+], V[(1-r)\mu_-]$ and $Cov[r\mu_+, (1-r)\mu_-]$ can be transformed into the following equations, respectively:

$$V[r\mu_+] = V_rV_+ + E^2_rV_+ + E^2_+V_r$$

$$\begin{aligned} V[(1-r)\mu_-] = \;& V[1-r]V_- + E[1-r]^2V_- + E^2_-V[1-r] \\ = \;& V_rV_- + (1-E_r)^2V_- + E^2_-V_r \end{aligned}$$

$$\begin{aligned} Cov[r\mu_+, (1-r)\mu_-] = \;& E[r(1-r)\mu_+\mu_-] - E[r\mu_+]E[(1-r)\mu_-)] \\ = \;& E[r(1-r)]E_+E_- - E_rE_+E[1-r]E_- \\ = \;& (E_r - E[r^2])E_+E_- - E_rE_+(1-E_r)E_- \\ = \;& E_rE_+E_- - E[r^2]E_+E_- - E_rE_+E_- + E^2_rE_+E_- \\ = \;& E_+E_-(E^2_r - E[r^2]) \\ = \;& -E_+E_-V_r \end{aligned}$$

As a result, from Eq. B.4, Eq. 8 can be derived as follows:

$$\begin{aligned} V[Y] = \;& V[r\mu_+] + V[(1-r)\mu_-] + 2Cov[r\mu_+, (1-r)\mu_-] + V[e] \\ = \;& V_rV_+ + E^2_rV_+ + E^2_+V_r + V_rV_- + (1-E_r)^2V_- + E^2_-V_r \\ & -2E_+E_-V_r + V[e] \\ = \;& (E_+ - E_-)^2V_r + E^2_rV_+ + (1-E_r)^2V_- + (V_+ + V_-)V_r + V[e] \end{aligned}$$

References

[1] Rodriguez-Esteban R, Jiang X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. BMC Med Genom 2017;10(1):1–10.
[2] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol 2014;15(12):1–21.
[3] Sun J, Nishiyama T, Shimizu K, Kadota K. Tcc: an r package for comparing tag count data with robust normalization strategies. BMC Bioinform 2013;14 (1):1–14.
[4] Robinson MD, McCarthy DJ, Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 2010. p. 139–40.
[5] Bentz AB, Thomas GW, Rusch DB, Rosvall KA. Tissue-specific expression profiles and positive selection analysis in the tree swallow (tachycineta bicolor) using a de novo transcriptome assembly. Sci Rep 2019;9(1):1–12.
[6] Xiao S-J, Zhang C, Zou Q, Ji Z-L. Tisged: a database for tissue-specific genes. Bioinformatics 2010;26(9):1273–5.
[7] Okada D, Okamoto Y, Io T, Oka M, Kobayashi D, Ito S, Yamada R, Ishii K, Ono K. Comparative study of transcriptome in the hearts isolated from mice, rats, and humans. Biomolecules 2022;12(6):859.
[8] Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philos Trans R Soc B: Biol Sci 2013;368(1620):20120362.
[9] Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eqtl studies. Trends Genet 2008;24(8):408–15.
[10] Ho JW, Stefani M, Dos Remedios CG, Charleston MA. Differential variability analysis of gene expression and its application to human diseases. Bioinformatics 2008;24(13):i390–8.
[11] Roberts AG, Catchpoole DR, Kennedy PJ. Identification of differentially distributed gene expression and distinct sets of cancer-related genes identified by changes in mean and variability. NAR Genomics Bioinform 2022;4(1):lqab124.
[12] Trojani A, Lodola M, Tedeschi A, Greco A, Di Camillo B, Sanavia T, Frustaci AM, Mazzucchelli M, Villa C, Boselli D, et al. Transcriptome analysis identified significant differences in gene expression variability between wm and igm-mgus bm b cell clones. Blood 2016;128(22):5089.
[13] Viñuela A, Brown AA, Buil A, Tsai P-C, Davies MN, Bell JT, Dermitzakis ET, Spector TD, Small KS. Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. Human Mol Genet 2018;27 (4):732–41.
[14] Ecker S, Pancaldi V, Rico D, Valencia A. Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. Genome Med 2015;7(1):1–12.
[15] Hulse AM, Cai JJ. Genetic variants contribute to gene expression variability in humans. Genetics 2013;193(1):95–108.
[16] de Jong TV, Moshkin YM, Guryev V. Gene expression variability: the other dimension in transcriptome analysis. Physiol Genomics 2019;51(5):145–58.
[17] Okada D, Zheng C, Cheng JH, Yamada R. Cell population-based framework of genetic epidemiology in the single-cell omics era. BioEssays 2022;44 (1):2100118.
[18] Fustero-Torre C, Jiménez-Santos MJ, García-Martín S, Carretero-Puche C, García-Jimeno L, Ivanchuk V, Di Domenico T, Gómez-López G, Al-Shahrour F. Beyondcell: targeting cancer therapeutic heterogeneity in single-cell rna-seq data. Genome Med 2021;13(1):1–15.
[19] Nathan A, Asgari S, Ishigaki K, Valencia C, Amariuta T, Luo Y, Beynor JI, Baglaenko Y, Suliman S, Price AL, et al. Single-cell EQTL models reveal dynamic t cell state dependence of disease loci. Nature 2022:1–9.
[20] Ota M, Nagafuchi Y, Hatano H, Ishigaki K, Terao C, Takeshima Y, Yanaoka H, Kobayashi S, Okubo M, Shirai H, et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell 2021;184 (11):3006–21.
[21] Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, Andersen S, Lu Q, Rowson A, Taylor TR, Clarke L, et al. Single-cell eqtl mapping identifies cell type–specific genetic control of autoimmune disease. Science 2022;376 (6589):eabf3041.
[22] Zhao J, Jaffe A, Li H, Lindenbaum O, Sefik E, Jackson R, Cheng X, Flavell RA, Kluger Y. Detection of differentially abundant cell subpopulations in scrna-seq data. Proc Nat Acad Sci 2021;118(22):e2100293118.
[23] Deng Z, Zhang M, Zhu T, Zhili N, Liu Z, Xiang R, Zhang W, Xu Y. Dynamic changes in peripheral blood lymphocyte subsets in adult patients with covid-19. Int J Infect Dis 2020;98:353–8.
[24] Orrù V, Steri M, Sole G, Sidore C, Virdis F, Dei M, Lai S, Zoledziewska M, Busonero F, Mulas A, et al. Genetic variants regulating immune cell levels in health and disease. Cell 2013;155(1):242–56.
[25] Okada D, Nakamura N, Setoh K, Kawaguchi T, Higasa K, Tabara Y, Matsuda F, Yamada R. Genome-wide association study of individual differences of human lymphocyte profiles using large-scale cytometry data. J Human Genet 2021;66 (6):557–67.
[26] Kanke M, Ng MMK, Connelly S, Singh M, Schaner M, Shanahan MT, Wolber EA, Beasley C, Lian G, Jain A, et al. Single-cell analysis reveals unexpected cellular changes and transposon expression signatures in the colonic epithelium of treatment-naïve adult crohn's disease patients. Cell Mol Gastroenterol Hepatol 2022;13(6):1717–40.
[27] Boland BS, He Z, Tsai MS, Olvera JG, Omilusik KD, Duong HG, Kim ES, Limary AE, Jin W, Milner JJ, et al. Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. Sci Immunol 2020;5(50):eabb4432.
[28] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc: Ser B (Methodological) 1995;57(1):289–300.
[29] Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. Genome Biol 2016;17(1):1–15.
[30] Okada D, Cheng JH, Zheng C, Yamada R. Data-driven comparison of multiple high-dimensional single-cell expression profiles. J Human Genet 2022;67 (4):215–21.
[31] Carter KM, Raich R, Finn WG, Hero III AO. Fine: Fisher information nonparametric embedding. IEEE Trans Pattern Anal Mach Intell 2009;31 (11):2093–8.
[32] Okada D, Yamada R. Decomposition of a set of distributions in extended exponential family form for distinguishing multiple oligo-dimensional marker expression profiles of single-cell populations and visualizing their dynamics. PloS one 2020;15(4):e0231250.
[33] Gingold JA, Coakley ES, Su J, Lee D-F, Lau Z, Zhou H, Felsenfeld DP, Schaniel C, Lemischka IR. Distribution analyzer, a methodology for identifying and clustering outlier conditions from single-cell distributions, and its application to a nanog reporter rnai screen. BMC Bioinform 2015;16(1):1–20.
[34] Grün D, Kester L, Van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nature Methods 2014;11(6):637–40.
[35] Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, et al. Accounting for technical noise in single-cell rna-seq experiments. Nature Methods 2013;10(11):1093–5.
[36] Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scrna-seq data. Genome Biol 2022;23(1):1–24.
[37] Van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene–disease predictions. Briefings Bioinform 2018;19(4):575–92.
[38] Okada D, Endo S, Matsuda H, Ogawa S, Taniguchi Y, Katsuta T, Watanabe T, Iwaisaki H. An intersection network based on combining snp coassociation and rna coexpression networks for feed utilization traits in japanese black cattle. J Anim Sci 2018;96(7):2553–66.
[39] Lee Y, Bogdanoff D, Wang Y, Hartoularos GC, Woo JM, Mowery CT, Nisonoff HM, Lee DS, Sun Y, Lee J, et al. Xyzeq: Spatially resolved single-cell rna sequencing reveals expression heterogeneity in the tumor microenvironment. Sci Adv 2021;7(17):eabg4755.
[40] Fernández AF, Bayón GF, Urdinguio RG, Toraño EG, García MG, Carella A, Petrus-Reurer S, Ferrero C, Martinez-Camblor P, Cubillo I, et al. H3k4me1 marks dna regions hypomethylated during aging in human stem and differentiated cells. Genome Res 2015;25(1):27–40.
[41] Talens RP, Christensen K, Putter H, Willemsen G, Christiansen L, Kremer D, Suchiman HED, Slagboom PE, Boomsma DI, Heijmans BT. Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs. Aging Cell 2012;11(4):694–703.
[42] Slieker RC, van Iterson M, Luijk R, Beekman M, Zhernakova DV, Moed MH, Mei H, Van Galen M, Deelen P, Bonder MJ, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. Genome Biol 2016;17 (1):1–13.
[43] Vershinina O, Bacalini M, Zaikin A, Franceschi C, Ivanchenko M. Disentangling age-dependent dna methylation: deterministic, stochastic, and nonlinear. Sci Rep 2021;11(1):1–12.