LTMG (Left truncated mixture Gaussian) based modeling of transcriptional regulatory heterogeneities in single cell RNA-seq data – a perspective from the kinetics of mRNA metabolism

Changlin Wan[1, 2], Wennan Chang[1, 2], Yu Zhang[1,3], Fenil Shah[4], Sha Cao[1,5], Xin Chen[6], Melissa Fishel[4] *, Qin Ma[7] *, Chi Zhang[1, 2] *

[1]Department of Medical and Molecular Genetics, [4]Department of Pediatrics and Herman B Wells Center for Pediatric Research, [5]Department of Biostatistics, Indiana University, School of Medicine, Indianapolis, IN,46202, USA.

[2]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN, 46202, USA

[3]Colleges of Computer Science and Technology, Jilin University, Changchun,130012, China

[6]Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China

[7]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, 57006, USA

*To whom correspondence should be addressed. +1 317-278-9625; Email: czhang87@iu.edu. Correspondence is also addressed to Melissa Fishel: +1 317-274-8810, Email: mfishel@iu.edu; Qin Ma. Tel: +1 605-688-6315; Email: qin.ma@sdstate.edu.

Abstract

A key challenge in statistical modeling of single cell RNA-seq (scRNA-seq) data is to assess the multimodality of single gene's expression with a simultaneous handling of the largely observed zero and low expressions. In this study, started from a mathematical derivation of the relationship between the mRNA abundance, transcriptional regulatory signals, and mRNA metabolism on a single cell level, we developed a left truncated mixture Gaussian (LTMG) distribution to accurately infer the modality and distribution of individual gene's expression profile in a scRNA-seq data, with modeling the dropout and low expressions as left censored data caused by a limited experimental resolution. We validated the LTMG model is with a much better goodness of fitting on an extensive data set, comparing to the statistical models commonly utilized in other scRNA-seq analysis methods. An LTMG based differential gene expression testing method, namely LTMG-DEG, is further developed. We experimentally validated a higher sensitivity and specificity of the LTMG-DEG comparing other differential gene expression testing methods. In addition, the assumption of the LTMG model, derived from our system biological derivation, is validated on several independent experimental data sets. A user friendly R package of LTMG and its downstream analysis is provided.

Introduction

Single cell RNA sequencing technique has been developed and applied to characterize transcriptomic heterogeneity of cells in a complex tissue or through a developmental process [1-3]. Comparing to the traditional transcriptomics data of bulk tissue samples that measures the averaged expression expressions across bulked cells, scRNA-seq quantifies gene-expression levels on individual cell level. Gene expression in a single cell is determined by (1) the transcriptional regulatory status in the current cell and (2) the metabolism rate of mRNA. In addition, gene expression level measured in a scRNA-seq data is also limited by the experimental resolution and results in a high number of zero observations, which are named as "dropout" events. Hence how to decipher the heterogeneous transcriptional regulatory signals encoded in the cells from the highly noisy background with "dropouts" forms a key challenge in the modeling of scRNA-seq data.

The scRNA-seq techniques are classified as individual-library or drop-seq based experiment by the library construction protocols [4-6]. The former class, represented by the C1-Fluidigm/Smart-seq2 protocol,

constructs a library for each individual cell and each library is sequenced with a high depth. On the other side, drop-seq techniques barcodes the transcripts from different cell samples and constructs a library by pooling the barcoded molecules. Differed by the library construct methods, the individual-library based technique is more utilized for a full assessment of the transcriptome of a medium number ($10\sim10^3$) of cells, while drop-seq measures gene expression with a less sequencing depth in a larger number ($10^3\sim10^6$) of cells. It is noteworthy that the experiment resolution for the cells measured by an individual-library based technique are relative similar due to all the libraries is fully sequenced with a fixed amount. On the contrary, the total count from each cell are more randomly measured from the pooled library in a Drop-Seq experiment. Multiple statistical distributions have been developed to model single gene's expression profile of the individual-library based data. Existing methods includes SCDE, MAST, Beta-Poisson and SC2P focusing on optimizing the best parametric form of the "dropout" proportions in the data [7-10]. Although all these models admitted the necessary of a multimodal assumption, none of the method links the observed gene expression level with their underlying regulators – the heterogeneous transcriptional regulatory states and mRNA transcription/degradation kinetics. Specifically, a more flexible assumption of the multimodality is needed to handle the genes regulated by multiple transcriptional regulators.

In this study, we first conducted a mathematical derivation of (1) the dependency between the transcriptional regulation inputs and the multi-modal distribution of mRNA abundance on a single cell level, and (2) the relationship of the suppressed regulatory state and un-fully degraded mRNA with the observed zero and low expressions in a scRNA-seq data. A novel statistical model namely left truncated mixture Gaussian (LTMG) distribution is further developed to accurately infer the transcriptomic heterogeneity of a single gene in an individual-library based scRNA-seq data, with modeling the "dropout" events and low expression as left censored data limited by the experimental resolution. LTMG-based differentially gene expression test, gene co-regulation, transcriptional burst, a,nd cell type inference analyses are further developed.

To validate the LTMG model and its downstream analysis, we compared the model with existing ones on 20 high quality scRNA-seq data sets collected from public domain. We also generated a scRNA-seq data with qPCR experiment to experimentally validate the modeling hypothesis and LTMG based differential gene expression test. Our key observations include (1) LTMG achieves a significant better goodness of fitting comparing to the Beta-Poisson, Zero-Inflated Gaussian (MAST) and Zero Inflated Mixture Gaussian models in more than 90% of analyze data sets, (2) LTMG more accurately characterizes cell type specifically expressed genes comparing to other models, (3) LTMG based differential gene expression analysis has a better identification accuracy than the existing methods. In addition, the system biology assumption of the LTMG model is validated by significant associations between the fitted LTMG parameters with experimentally measured mRNA kinetics. The full computational approach including LTMG fitting and downstream analysis are released through an open source R package.

Results

Mathematical modeling of gene expression in a single cell from a perspective of transcriptional regulation

A gene's expression in a mammalian cell is regulated by the interactions between its DNA molecule and a collection of transcriptional regulatory inputs (TRIs) including (1) transcriptional regulatory factors (TFs) (cis-regulation), (2) miRNA or lncRNA, (3) enhancer and super-enhancer, and (4) epi-genetic regulatory signals such histone modification, chromatin folding and DNA methylation. For a gene with P possible transcriptional regulation inputs $TRI_i, i = 1, \dots, P$, the probability of its promoter being bound by an RNA polymerase $\mathrm{P_b}$, which is proportional to the transcriptional rate, can be modeled by a Michealis Menten model [11]:

$$\mathrm{P_b} = \frac{R_0 + \frac{R_1[TRI_1]}{K_1} + \cdots \frac{R_N[TRI_P]}{K_N} + \frac{R_{1,2}[TRI_1][TRI_2]}{K_{1,2}} + \cdots + \frac{R_{1,\dots,N}[TRI_1][TRI_2]\dots[TRI_P]}{K_{1,2,\dots,P}}}{1 + \frac{[TRI_1]}{K_1} + \cdots \frac{[TRI_P]}{K_N} + \frac{[TRI_1][TRI_2]}{K_{1,2}} + \cdots + \frac{[TRI_1][TRI_2]\dots[TRI_P]}{K_{1,2,\dots,P}}} = \frac{\sum_{\Omega\in M\{1\dots P\}} \frac{R_\Omega}{K_\Omega} \prod_{i\in\Omega}[TRI_i]}{\sum_{\Omega\in M\{1\dots P\}} \frac{1}{K_\Omega} \prod_{i\in\Omega}[TRI_i]} \quad (*)$$

where $R_i, [TRI_i], K_i$ denote production rate, concentration and kinetic parameters associated with the ith TRI; $M\{1 \dots P\}$ is the power set of $\{1 \dots P\}$, Ω denotes an element in $M\{1 \dots P\}$; R_Ω, K_Ω denote the production rate and kinetic parameters associated with the subset of TRIs in Ω. Specifically, we call each Ω as a transcriptional regulatory state (TRS), which is determined by the combination of its TRIs, and reflected by the observed expression in a single cell. Noting that in a single cell the state of each TRI can be rationally simplified to either bound ON or OFF to the DNA molecule, thus the TRI_i is a Boolean variable and the equation (*) is a step function with at most $|M\{1 \dots P\}| = 2^{\mathrm{P}}$ plateau levels:

$$\mathrm{P_b}(Current\ TRS = \{TRI_i, i \in \Omega\}) = \mathrm{P_b}\left(\left\{[TRI_i] \gg 0, [TRI_j] = 0 \middle| \ i \in \mathcal{M}, j \notin \Omega\right\}\right) = R_\Omega$$

For a mammalian cell, the total number of possible TRS can be substantially large due the diverse number and types of TRS, especially the epi-genetic regulators [12,13]. However, the number of TRSs of a gene in a single cell RNA-seq experiment is always much smaller, majorly due to (i) the phenotypic diversity of the cells measured in one experiment is relatively small, and (ii) master repressors such as chromatin folding or certain TFs that can totally suppress the gene's expression.

Denote the expression level of a gene X as $\mathrm{G^X(t)}$, the probability of X's promoter being bound by a RNA polymerase at time t as $\mathrm{P_b^X(t)}$, and its transcriptional and mRNA degradation rate at time t as $\mathrm{T^X}(t)$ and $\mathrm{D^X}(t)$. Noting that the transcriptional rate is proportional to the $\mathrm{P_b}$ in a single cell, and assuming $\mathrm{D^X}(t) \propto \mathrm{G^X(t)}$, we have

$$\mathrm{G^X(t)} \propto \int \left(\mathrm{P_b^X(t)} - \mathrm{C^X} \cdot \mathrm{G^X(t)}\right) dt$$

$$\mathrm{P_b^X(t)} = \left\{R_{\Omega_\mathrm{i}}^\mathrm{X} \middle| i = 1 \dots K \text{ corresponds to each TRS}\right\}$$

, where $\mathrm{C^X}$ is the kinetic parameters of the mRNA degradation of X in the current cell, and R_{Ω_i} is the RNA polymerase binding probability corresponds to a TRS, by which the X's expression level under a long term regulation of the ith TRS, denoted as $\mathrm{G_{\Omega_i}^X} \triangleq \lim_{T\to\infty} \int_0^T \left(R_{\Omega_\mathrm{i}}^\mathrm{X} - \mathrm{C^X} \cdot \mathrm{G^X(t)}\right) dt = \frac{R_{\Omega_\mathrm{i}}^\mathrm{X}}{\mathrm{C^X}}$, is also a constant. For a scRNA-seq experiment, we call the TRS of X with $\mathrm{G_{\Omega_i}^X}$ can be effectively observed by the experiment as an Active Regulation (AR) TRS and the TRS with $\mathrm{G_{\Omega_i}^X}$ below the experimental resolution as a Suppressed Regulation (SR) TRS.

Denote $\tilde{\mathrm{x}}_\mathrm{j}, \mathrm{j} = 1 \dots \mathrm{N}$ as the normalized gene expression level (such as CPM or TPM) of gene X in a scRNA-seq experiment with individual library constructed for N cells and measured with a large sequencing amount. With the above considerations of mRNA dynamics and experimental resolution, and assuming the switch among the TRSs are relatively stable in through the experimental time, we can derive a natural relationship between the repertoire of the TRSs of a single gene and its gene expression profile observed in a scRNA-seq experiment (Figure 1). Specifically, (i) $\tilde{\mathrm{x}}_\mathrm{j}$ can unbiasedly reflect the $\mathrm{G_{\Omega_i}^X}$ of the AR TRS of X in cell j with a log-Gaussian error when $\tilde{\mathrm{x}}_\mathrm{j}$ is large; (ii) an observed low (non-zero) level of $\tilde{\mathrm{x}}_\mathrm{j}$ can be caused by multiple reasons including true zero expression with a sequencing error, SR TRS of X in cell j, and un-fully degraded mRNA after the switch from an AR TRS to a SR TRS; and (iii) the observed zero value of $\tilde{\mathrm{x}}_\mathrm{j}$ can be true zero expression or a SR TRS undetected by the experiment resolution. To overcome the undistinguishable TRS and dynamics errors of low values of $\tilde{\mathrm{x}}_\mathrm{j}$, we introduce a latent threshold $\mathrm{Z_{cut}^X}$ depending on both SR TRSs and degradation of X. When $\log(\tilde{\mathrm{x}}_\mathrm{j}) > \mathrm{Z_{cut}^X}$, $\tilde{\mathrm{x}}_\mathrm{j}$ can reflect $\mathrm{G_{\Omega_i}^X}$ with a log-Gaussian error, while the TRS of X cannot be reliably inferred when $\log(\tilde{\mathrm{x}}_\mathrm{j}) \leq \mathrm{Z_{cut}^X}$. Figure 1 shows the relationship between the TRS of gene, $\log(\tilde{\mathrm{x}}_\mathrm{j})$, and the threshold $\mathrm{Z_{cut}^X}$. For the $\log(\tilde{\mathrm{x}}_\mathrm{j}) > \mathrm{Z_{cut}^X}$, the values follow a mixture Gaussian distribution truncated by the left of $\mathrm{Z_{cut}^X}$, with each Gaussian peak corresponds to one AR TRS of X. For the $\log(\tilde{\mathrm{x}}_\mathrm{j}) \leq \mathrm{Z_{cut}^X}$, the values should be from one or multiple SR peaks or true zeros but cannot be inferred. The rationale of such a definition is that the outcome of all the TRS with mean expression value lower than Zcut cannot be effective observed with the resolution of the scRNA-seq data set, and be distinguished from 0 observations, hence are considered as one SR peak. As shown in Figure 1, our derivation splits the observed gene expression in a scRNA-seq data into three parts: (i) the generally defined dropout events (zero expressions) in scRNA-seq data, (ii) low expressions correspond to a SR TRS, and (iii) expression of AR TRSs.

It is noteworthy that Z_{cut}^{X} is not equal to the boundary between the AR TRS with the smallest $G_{\Omega_i}^{X}$ and the SR TRS for most cases. For an accurate distinguishing of TRS from observed gene expression profile, we categorize the observed low expression side into six parts, as shown in the zoom-in part in Figure 1: (1) True Non-Expression, (2) SR TRS & un-fully degraded mRNA, (6) True AR expression, (4) Undetected True Expression, (5) Type II error of True expression, and (3) Type I error of SR TRS & un-fully degraded mRNA, with detailed definition given in Table 1. A key question here is to accurately identify the parameters of each AR TRS from the observed gene expression profile, i.e. infer the distribution parameters correspond to parts (1), (2) and (3) and parts (6), (4) and (5).

Table 1.

Observed gene expression\ TRS	Suppressed regulation	Activated Regulation
Below Z_{cut}^{X}	(1) True Non-Expression	(4) Undetected True Expression
Above Z_{cut}^{X}, below the boundary between the SR TRS and most left AR TRS	(2) SR TRS & un-fully degraded mRNA	(5) Type II error of True expression
Above the boundary between the SR TRS and most left AR TRS	(3) Type I error of SR TRS & un-fully degraded mRNA	(6) True AR expression

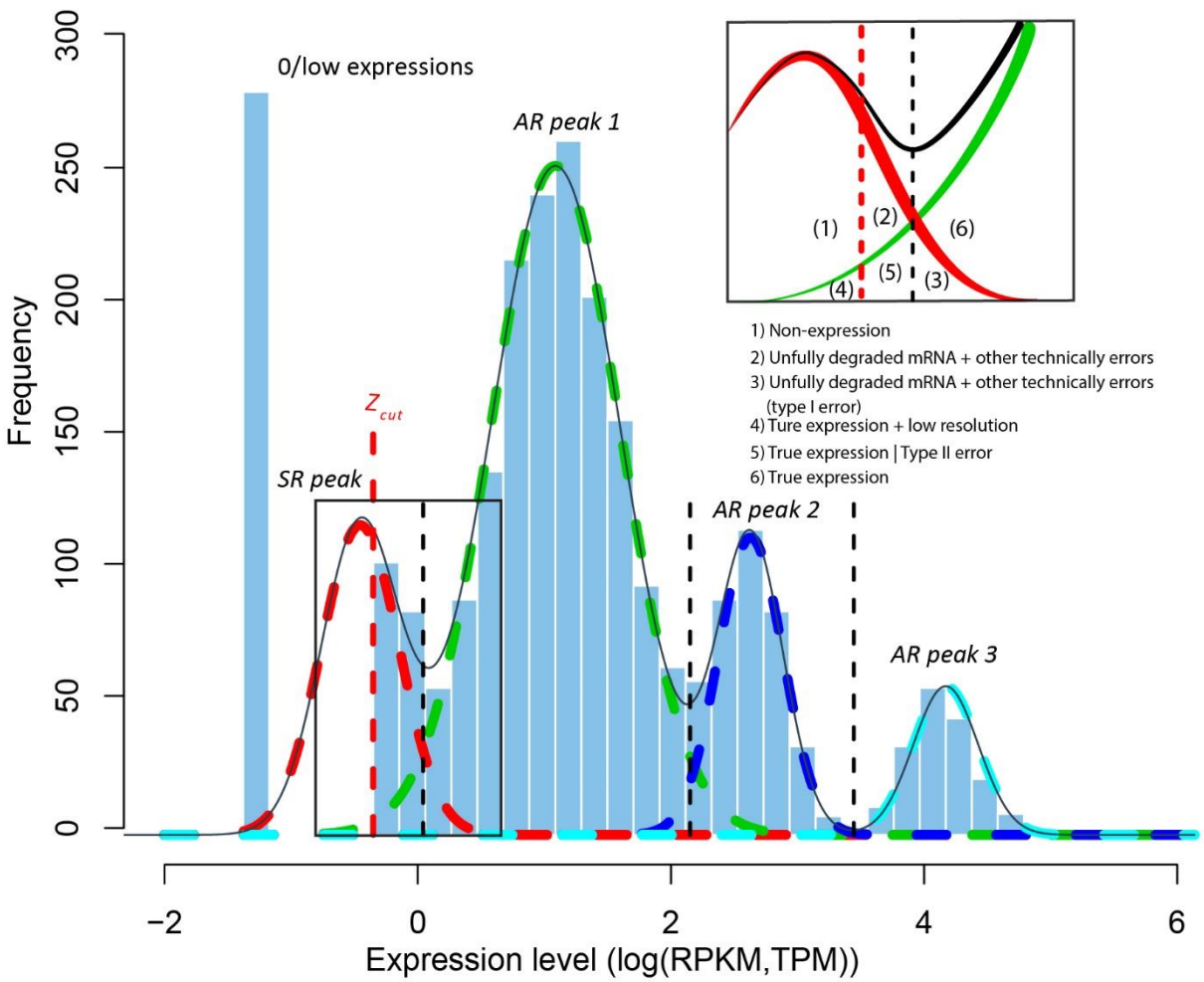

Figure 1. The relationship between observed genes expression level, the gene's SR and AR TRSs, and the experiment resolution threshold Z_{cut}^{X}. The light blue histogram illustrates the distribution of the log normalized gene expression (RPKM or TPM) of one gene in a scRNA-seq data. The four dash curves represent the distribution of the gene expression regulated by one SR and three AR TRSs of the gene. Z_{cut}^{X} is shown by the

red dash line. The frequency of the gene expression below Z_{cut}^{X} is shown by the most left bar in the histogram. The figure on top right is a zoom in version of the low expression part.

Benchmarks from previous studies

Our derivation suggests that the identifiable AR TRSs of X form a truncated mixture distribution on the right of Z_{cut}^{X}. Multiple established models have confirmed the necessary of using a multi-modal distribution to characterize single gene's expression profile in scRNA-seq data. Table 2 listed the commonly used methods and models for scRNA-seq gene expression modeling and differential gene expression tests. In detail, SCDE uses a fixed Poisson distribution to model the drop out events and a negative binomial distribution to model the expression part of a gene [7]. MAST considers zero observations as dropouts and the rest as following a log-normal distribution [8]. Trung Nghia Vu et al considers the underlying transcriptional heterogeneity can be fitted by a Beta distribution and utilized a zero inflated Beta-Poisson distribution for gene expression modeling. SC2P assumes all the genes in one cell share a same zero inflated negative binomial distribution for the dropout events [10]. It is noteworthy that both SCDE and SC2P aims to derive a common parameter set to model the dropouts for all the genes in one or multiple cell samples; MAST only considers the zero parts in (1) and (4) in Figure 1 as dropouts; and Beta-Poisson considers the centers of G^{X} through the cells follow a Beta distribution. However, none of the model links the observed expression profile with the underlying regulators of the gene expression—heterogeneity of TRSs and mRNA metabolism.

Based on our derivation from a perspective of the dynamics of TRSs, as shown in Figure 1, expressions in part (1) and (4) are unreliable for inference of the distributions, part (2) and (6) are the gene expression of SR TRSs or un-fully degraded mRNAs, the boundary of which depends on the transcriptional dynamics and mRNA degradation of the gene, and (3) and (5) are from an AR TRS of the gene. Hence, theoretically, it is necessary to have a gene-wise consideration of Z_{cut}^{X}, SR and AR TRSs for an accurate modeling of individual gene expression profile in scRNA-seq data.

Table 2.

	#parameters	Consideration of Dropout	Consideration of expression	Data
SCDE	3	Fixed Poisson	Negative Binomial	Counts
MAST	3	Zero inflation	Log Gaussian	TPM/CPM/…
Beta-Poisson	4	0 inflated beta-Poisson	beta-Poisson	TPM/CPM/…
SC2P	5	Sample wise Negative Binomial	Log Gaussian	Counts
LTMG-LR	5	gene wise censored data	Log Gaussian	TPM/CPM/…
LTMG	Depends on fitting	gene wise censored data	Log Mixture Gaussian	TPM/CPM/…

Left Truncated Mixture Gaussian (LTMG) distribution for gene expression modeling

To accurately model the gene expression profile of scRNA-seq data, we developed a mixed Gaussian model with left truncation assumption namely LTMG to fit the log transformed normalized gene expression measures such as TPM, CPM or RPKM [refs]. The fundamental idea here is to treat the large amount of observed zero and low expressions below Z_{cut}^{X} as left censored data in fitting a mixture Gaussian model of each gene's expression profile. The rationality of this assumption is that the zero and low expressions below the lowest experimental resolution are naturally considered as missing data that from a SR TRS.

Denote the observed log-transformed normalized expression level of gene X over N cells as $X = (x_1, x_2, \dots, x_N)$. We assume that $x \in X$ follows a mixture Gaussian distribution with K Gaussian peaks corresponding to 1 SR and K-1 AR TRS, and the density function of X can be written as:

$$p(X|\Theta) = \prod_{j=1}^{N} p(x_j|\Theta) = \prod_{j=1}^{N}\sum_{i=1}^{K} a_i p_i(x_j|\theta_i) = \prod_{j=1}^{N}\sum_{i=1}^{K} a_i \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}} = L(\Theta|X) \qquad (*)$$

, where parameters $\Theta = \{a_i, u_i\ \sigma_i \mid i = 1 \dots K\}$ and $a_i, u_i\ and\ \sigma_i$ are the proportion, mean and standard deviation of the Gaussian peak respectively corresponds to each TRS. We introduce a parameter Z_{cut}^{X} and consider the log transformed zero and low expression values smaller than Z_{cut} as left censored data. With the left truncation assumption, X are split into M reliably measured expressions ($x_j > Z_{cut}^{X}$) and N-M left-censored gene expressions ($x_j \le Z_{cut}^{X}$). Θ can be estimated by using an EM algorithm with given Z_{cut}^{X} and K (see details in

Methods). In this paper, Z_{cut}^{X} is set as logarithm of the minimal non-zero expression value of each gene. The EM algorithm is conducted for K = 1, 2, 3,… and the K gives the smallest value of the Bayesian Information Criterion (BIC) is selected (Methods).

Model evaluation on large scale scRNA-seq data sets from public domain

In order to conduct an unbiased evaluation of our LTMG versus the other models, we collected human scRNA-seq data sets conceived by the NCBI-GEO database dated from January 2016 to April 2018 that were generated by using Smart-seq or Smart-seq2 protocols with a sample size between 100 to 6000 cells, consists 20 data sets totaling 49936 cell samples (majorly distributed from 2000 to 6000, see more details in Method). We split and classified the collected data sets into three categories, namely, (i) 51 Pure Condition (PC), (ii) 49 Cell Cluster (CC) and (iii) 78 Complete Data (CD) sets. Specifically, the PC group are consisted by the data sets of a pure cell type under a tightly controlled condition, such as the experiment of a cell line under a fixed condition, or cells classified by surface markers; the CC data sets are of computational predicted cell clusters from large scale data sets, such as predicted cancer or immune sub cell types in a cancer tissue scRNA-seq data; and the CD group are just complete data sets (see more details in Methods). It is noteworthy that the sample size requirement in our data selection procedure guarantees a relatively balanced sample size among the three groups of data.

We applied LTMG and zero-inflated Mixture Gaussian (ZIMG), and two commonly utilized models of log-normalized scRNA-seq data, namely MAST and Beta-Poisson (BPSC) model to fit the expression profile of each gene in the three groups of data sets. Kolmogorov Statistics (KS) [14] was applied to evaluate the goodness of fitting of each gene. Measured with the KS statistics, and a Goodness of Fitting (GF) score defined by the average level of KS of all genes in a data set (see Methods), our analysis clearly suggested that LTMG is with a better goodness of fitting comparing to BPSC and MAST in all the analyzed data sets, outperformed ZIMG in 143 out of the 178 data sets, and tied with ZIMG in the rest of data sets (figure 2A, 2C). In addition, we found that the LTMG model is generally with a smaller number of outliers of poor fitting through all the three groups of data, as illustrated in figure 2B, suggests that LTMG is with a more robust fitting comparing to the others. Due to the BPSC is with a much higher KS comparing to the others, we further focus on a more detailed comparison of LTMG with ZIMG and MAST, on the genes fitted with different number of Gaussian peaks. Within the CC and CD groups, LTMG consistently outperformed ZIMG (120/127) and MAST (127/127). In the PC group, LTMG outperformed the MAST model in all the tested data (51/51), outperformed ZIMG (42/51) for the genes fitted with more than two Gaussian peaks, and have a similar performance as ZIMG (23/51), specifically for the genes that are fitted with one and two Gaussian peaks. A possible reason of the less significant performance of LTMG in the PC data is that the sample size of the PC group is generally small (~115 cells) compared to CC (~ 388 cells) and CD (~622 cells), which cause the half bell shape of the SR peak is not significantly different to a full Gaussian peak.

Our analysis suggested that the average proportion of genes fitted with one, two, and more than two Gaussian peaks are 42.5%, 44.9% and 12.6% in the PC, 16.6%, 65.7% and 17.6% in the CC, and 25.4%, 51.5% and 23.1% in the CD data sets, respectively (Figure 2C). We define an LTMG model of two Gaussian peaks, with the mean of one peak smaller than Z_{cut}^{X}, and the mean of the other peak larger than Z_{cut}^{X}, as a LTMG-2LR distribution (see more details in Method). We further checked the genes fitted by two peaks, which take the majority of the genes in each data group. On average, 83.1% of such genes in the PC, 96.2% in the CC, and 95.6% in the CD group are fitted with a LTMG-2LR distribution, suggesting most of the genes are with one SR peak and one AR peak in the CC and CD data.

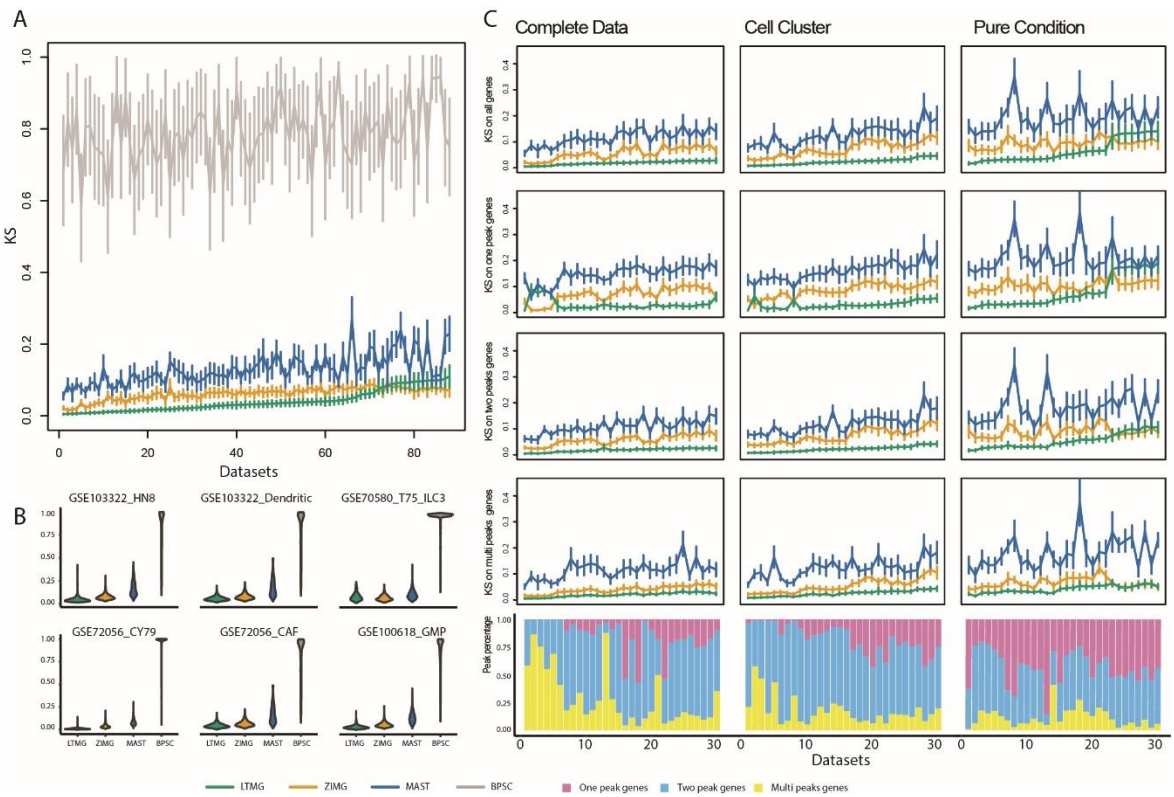

Figure 2. Performance comparison of LTMG and stand-of-art models. Green, orange, blue and gray represents LTMG (Left Truncated Mixture Gaussian), ZIMG (Zero Inflated Mixture Gaussian), MAST (Zero Inflated Gaussian) and BPSC (Beta Poisson). (A) 4 model comparison, horizontal line is the mean of KS value in each dataset, vertical line represents the standard deviation. (B) Violin plot of KS value of selected example datasets, 2 for each group. (C) Detailed three model comparison on genes of different peak and datasets of different group. Horizontal line represents the mean of KS fitting of value in that group of genes and vertical line is the standard deviation accordingly. Stocked histogram illustrates the percentage distribution of genes of different peaks in different datasets.

Association between the fitted TRSs and mRNA metabolism kinetics

Our mathematical derivation suggested the necessary of using a mixture model for an inference of SR/AR expression for scRNA-seq data analysis. It is noteworthy that the uncensored low expression part ((2) and (5) in Figure 1), including both SR TRS & un-fully degraded mRNA and Type II error of True expression are generally seen in all the analyzed data sets, which on average take 27.9%, 16.3% and 14.5% of non-zero values in the PC, CC and CD data. We hypothesized that one major contributor of the un-censored low expression is the un-fully degraded mRNA under a SR TRS (Figure 3A), which should be excluded from the inference of AR TRS.

To validate our hypothesis, we collected a data set of experimentally measured mRNA kinetics of mouse fibroblast cells [15] and tested the correlations between the mRNA half-life and the LTMG fitted parameters of each gene, in three scRNA-seq data set (GSE99235, GSE98816 and GSE87038) of mouse fibroblast cells [16,17]. To the best of our knowledge, this is the only data set and cell type with both whole genome level kinetics of mRNA metabolism and scRNA-seq data available in the public domain. Specifically, a similar pattern of positive correlations between (i) the proportions of uncensored observations in the SR peak, defined by $\frac{(2)+(5)}{(1)+(2)+(4)+(5)}$ in Figure 1, and (ii) mRNA half-life, has been observed in all the three data sets (Figure 3B), suggesting that the genes with more uncensored expressions under a SR peak are with relatively slower mRNA degradation rate (longer mRNA half-life).

Since the uncensored observations in the SR peak are consisted by un-fully degraded mRNA (2) and Type II error of an AR expression (5), genes with a higher mean of the AR peak are with a less proportion of

type II errors (Figure 3A). To control the impact of the type II error of an AR expression in the correlation analysis, we split genes into groups by the means of their AR peak, and correlate mRNA half-life with the proportion of uncensored observations in the SR peak in each gene group (see details in Methods). Significant positive correlations ($p<0.1$) in the gene groups with highly expressed AR peak, and a significant positive dependency between the correlation and mean of AR peaks, were consistently identified in all the three data sets (Figure 3C). Figure 3D illustrate the detailed proportion of uncensored observations in the SR peak and mRNA half-life of the genes with top highly expressed AR peak, in the three data sets.

Our analysis clearly suggested that un-fully degraded mRNA forms a significant source of the uncensored observations in the SR peak, hence distinguishing the expression led by un-fully degraded mRNA from true AR expressions is critical for an accurate modeling of scRNA-seq data.

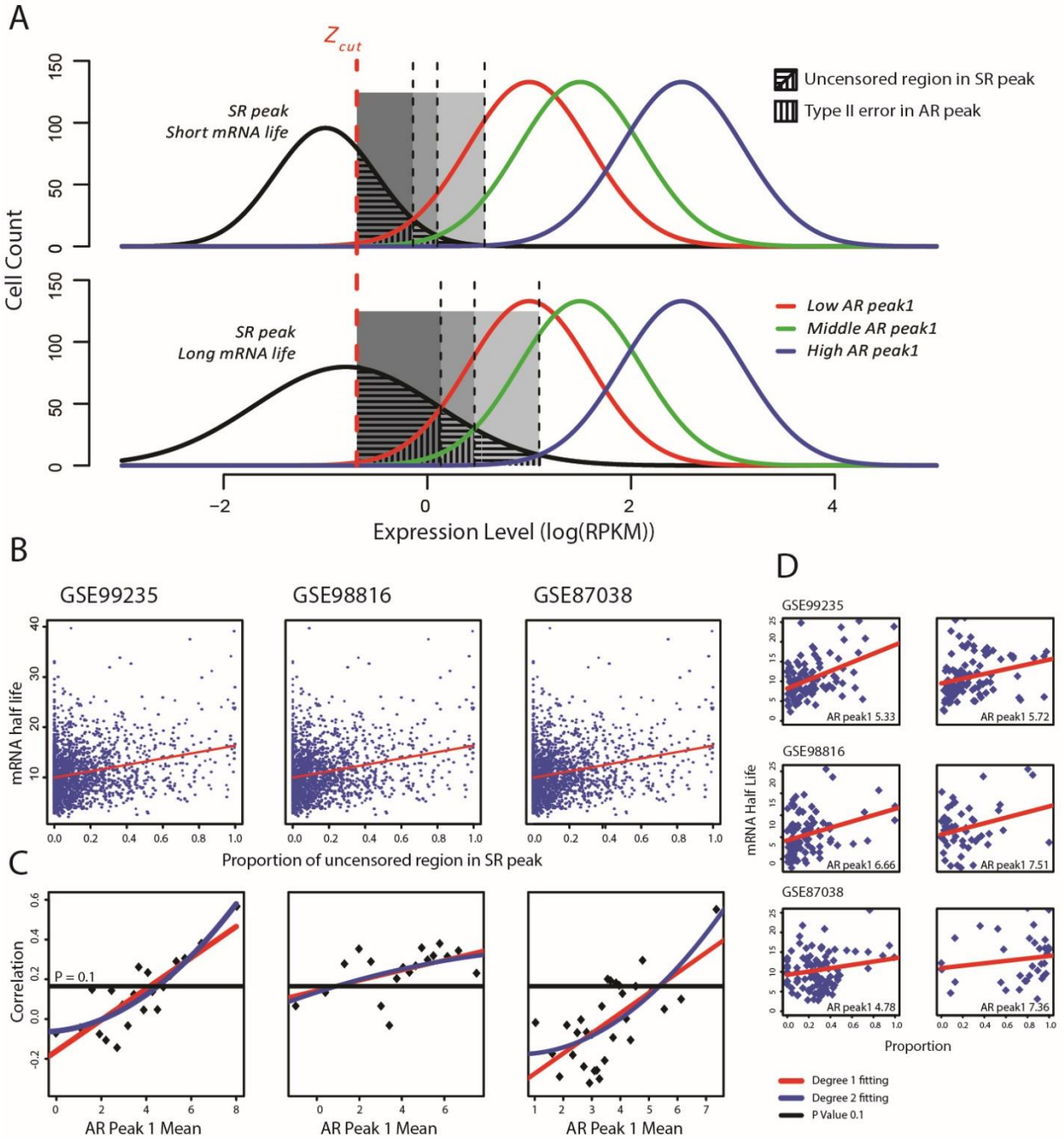

Figure 3. Association between the SR peak and mRNA degradation. (A) Schematic of uncensored region of genes with different SR peak and influences from different AR peak1. Genes with longer mRNA life tend to have a larger uncensored region. Lower AR peak1 is more likely to introduce a bigger Type II error. (B) Scatter plot of uncensored region and mRNA half-life in three different datasets. Red line is the degree 1 fitting. (C) Scatter plot of correlation value in different AR peak1 Mean. Red line is degree 1 fitting, blue line is degree 2 fitting, and black line is the correlation threshold when P value is equal to 0.1. (D) Example scatter plot of uncensored region and mRNA half-life of genes in high AR peak1 level.

Method validation on a scRNA-seq data sets of human cancer, stromal and immune cells

In order to check the biological implication of the LTMG model in large scRNA-seq data analysis, we applied the model on a head and neck cancer (HNSC) data set (GSE103322) consists of 5902 cells of 9 cell types names B cell, T cell, Myocyte, Macrophage, Endothelial, Dendritic and Mast cell, with uniquely expressed makers genes well-defined for each cell type [1]. Unsurprisingly, we commonly observed the marker genes were fitted with one SR peak and one or more AR peaks. We developed an E (enrichment) score to measure the cell type enrichment in different parts of the expression profile of each gene. A greater E score suggests certain part of a gene's expression profile is more enriched by one cell type (see methods).

Our analysis identified that, for all the cell type marker genes, the AR peak is with a much higher E score comparing to the uncensored part in the SR peak, suggesting a better specificity in identification of true AR expressed genes by using the LTMG model, comparing to the "only-0" or "fixed-Poisson" consideration for dropout events by MAST or SCDE. Figure 4A illustrates the E score of the AR peak (AR), uncensored part in the SR peak (SR), and total non-zero expressed genes (EXP), of 16 cell markers in the data of T and fibroblast cells. All the 16 markers show highest E score of their corresponding cell type in the AR peak and a much lower E score in uncensored SR part. Figure 5B and 5C show the fitted peak distribution of a T cell marker GZMK and a fibroblast marker COL6A3. We further examined the distribution of the AR expression and uncensored SR expression of these two genes in the 2D-tSNE distribution of the cells derived by the complete data (Figure 5D). We have seen that the T cells with the AR expressions or uncensored SR expressions of GZMK were clearly separated, while cells of other cell types with uncensored SR expression are randomly distributed over the tSNE map (Figure 5E). Consider GZMK encodes a cytotoxicity molecule, and with the distribution of CD8A and CD8B in T cells, we speculate the two T cell groups, separated by the AR and uncensored SR expression of GZMK, correspond to high cytotoxic and exhausted CD8+ T cells in the HNSC microenvironment [18-20]. Similarly, the fibroblast cells with an AR or an uncensored SR expression of COL6A3 were differentially distributed as two sub fibroblast types, the cells of other cell types with uncensored SR expression are randomly distributed, and the cells of other cell types with an AR expression of COL6A3 tend to be certain T or malignant cells. These observations suggest that the AR expression of key marker genes, derived by our LTMG model, are informative for inference of cell type or sub cell type populations.

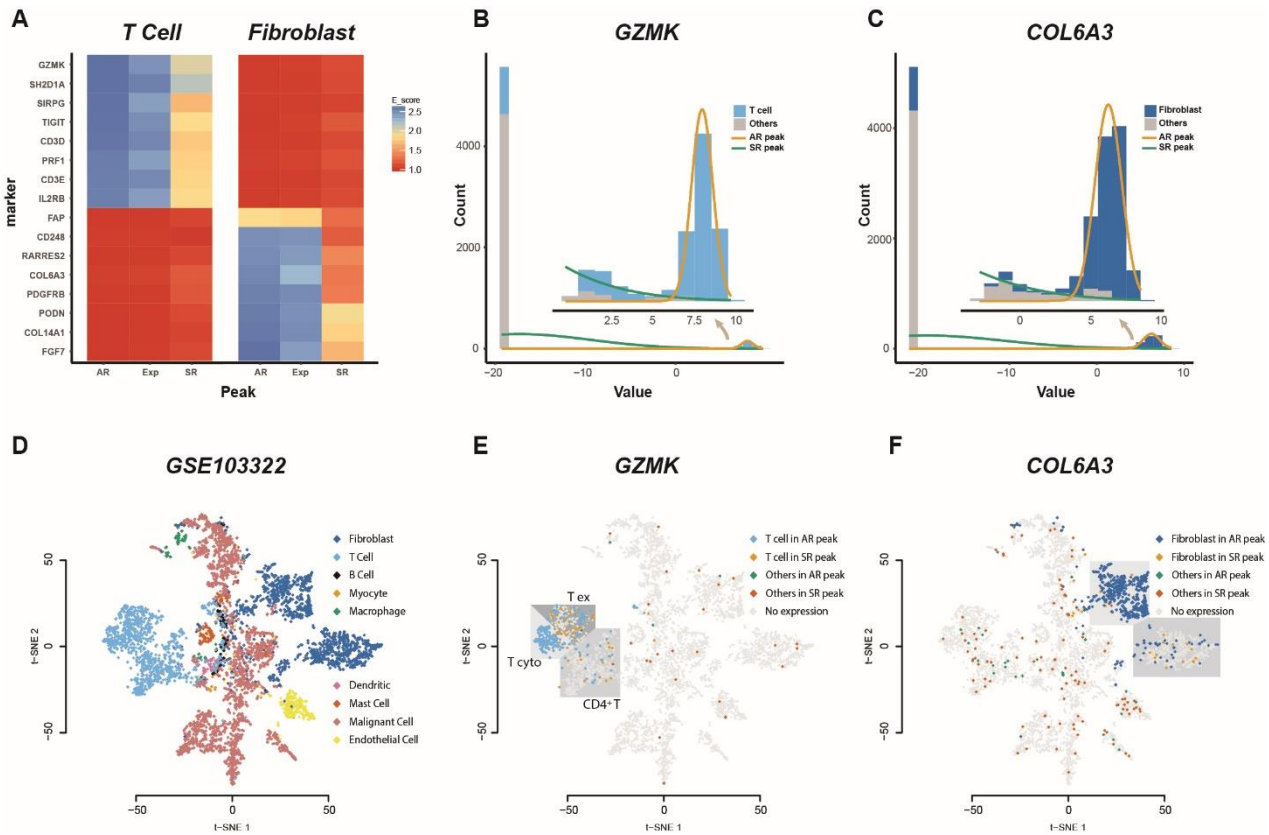

Figure 4. Distribution of AR and uncensored SR expression of cell type markers through different cell types. A) Heat map of T cell and fibroblast enrichment information across T cell and fibroblast markers. B, C) Cell distribution and peak fitting in GZMK and COL6A3. Light blue region is T cell, dark blue is Fibroblast and gray represents other cells. D) t-SNE plot of cells in GSE103322 dataset. E) Detailed T cells and other cells (Not T cell) distribution on GZMK on t-SNE plot. F) Detailed Fibroblast and other cells (Not Fibroblast) distribution on COL6A3 on t-SNE plot.

LTMG based differential gene expression analysis

With the LTMG model and its link with TRSs, we define *a gene is differentially expressed between the cells of two conditions, if at least one TRS (either SR or AR) of the gene is with a significant different representing level in one condition versus the other*. With this definition, we formulate the computation of a differentially expressed gene (DEGs) in a scRNA-seq data by two steps – (1) identifying all the TRSs of the gene through the cells of the to-be-tested conditions, and (2) a statistical test of if one TRS is with a significantly higher representation rate in one of the to-be-tested conditions.

Our comprehensive analysis revealed that on average more than 83.8% genes in the PC and CC groups are fitted with one and two peaks, which can be well fitted by a LTMG-2LR model. In order to (1) guarantee the mathematical rigorousness, (2) maximize the statistical power and (3) ensure the implementation of a GLM model for a multiple or crossed condition test, we developed a statistical testing approach by using the LTMG-2LR model for the genes with one or two fitted TRSs, and the LTMG model for the genes with more than two TRSs (see detailed derivations in Methods). For a given gene X in a scRNA-seq data of J conditions, denote $X_j = \{x_i^j, i = 1 \dots N_j\}, j = 1 \dots J$ as its expression profile in the N_j cells of the jth conditions. The following pseudo codes illustrate our differential gene expression analysis approach, namely LTMG-DEG, with detailed formulations of each step given in Method.

Fit a LTMG model for X′s expression profile of all conditions $\{X_1, \ldots X_J\}$ with $\Theta = \{a_i, u_i, \sigma_i | i = 1 \ldots K\}$,
where K is the number of peaks of the best fitting

If $K \leq 2$

Fit an LTMG_2LR model with $\Theta_j = \{a_{j1}, u_0, u_{j1}, \sigma_0, \sigma_{j1}\}$for each $X_j \in \{X_1, \ldots X_J\}$;

Compute the likelihood that x_i^j belongs to the SR or AR peak by using Θ_j;

Compute the p values of if one conditon is significantly associated with a different represntation rate of the SR or AR peak by using a GLM based approach;

Compute the p values of if the mean of the AR peak of one condition is significantly different comparing to other by using a GLM based approach;

Else

Compute the likelihood that x_i^j belongs to each of the K peaks;

Compute the p value of if samples from one condition are significantly enriched by one Gaussian peak, by using a hypergeometric test;

X is considered as differentially expressed if any of the p value is small than the significance cutoff

Experimental validation of LTMG-DEG

To experimentally validate the LTMG model based differential gene expression test, we generated a scRNA-seq data set consisting 180 patient-derived pancreatic Pa03c cancer cells collected under four experimental conditions and used qPCR to evaluate the different DEGs identified by our LTMG-DEG and the MAST methods (see more details in Methods). *APEX1* is a multifunctional protein that interact with multiple transcriptional factors (TFs) to regulate the genes involvement in response to DNA damage, hypoxia and oxidative stress [21]. Our previous study identified significant roles of APEX1 in the regulation of Pa03c cell's response to microenvironmental stresses [22]. In this experiment, we collected transcriptomic profiles of 40 APEX1-Knock Down (KD) Pa03c cells and 40 cells of scrambled control under a hypoxia condition, and 46 APEX1-KD cells and 46 control cells under a normoxia condition. MAST and LTMG-DEG were applied to identify DEGs among the four conditions. By using LTMG-DEG, we identified 448 up and 1397 down regulated genes in APEX1-KD vs control under hypoxia, and 471 up and 992 down regulated genes under normoxia ($p<0.01$); while MAST identified 282 and 521 up regulated and 397 and 607 down regulated genes, under hypoxia and normoxia conditions, respectively ($p<0.01$). One central function of APEX1 is regulation of transcription activation through interacting with multiple TFs related to stress response, such as HIF1A and STAT3 [23]. Considering a loss of APEX1 may affect the DNA binding of such TFs, we speculate it is rational to see more down regulated genes in siAPE1 cells versus controls. Distribution of the LTMG-DEG and MAST identified DEGs are shown in Figure 5A.

Pathway enrichment and distribution of the DEGs identified by the two methods were further examined. Comparing to MAST, the down regulated genes specifically identified by LTMG-DEG more significantly enrich the glycolysis, TCA cycle and respiration chain, apoptosis, and lipid metabolism pathways, and the genes regulated by HIF1A and STAT3 in the hypoxia data (Figure 5B). Noting that APEX1 directly interact the Hypoxia Induce Factor 1A (HIF1A) and STAT3, and regulate oxidative stress response, glucose and lipid metabolism, and relevant mitochondrial functions, the results suggested the LTMG-DEG is a much higher sensitive comparing to MAST.

We utilized qPCR to experimentally to validate 12 selected DEGs with top differences of significance identified by the two methods, and seven DEGs commonly identified by both methods (see details in Method). Specifically, in the comparison of APEX1-KD vs control under the hypoxia condition, (1) nine genes namely STAT3, CREM, SP1, USP3, CDS1, ACTR1A, PARP4, TMEM144, and MNAT1 were identified as down regulated genes by LTMG-DEG and un-differentially expressed by MAST; (2) three genes namely SEM1, PARPBP and RAP2C were identified as up-regulated genes by MAST and un-differentially expressed by LTMG-DEG; (3) two genes namely MKI67 and TMPO were identified as up-regulated genes by both methods; and (4) five genes namely JUNB, LYPLAL1, PRDM1, PGK1 and SLC2A3 were identified as down regulated genes by both methods. The qPCR results (Figure 5X) demonstrated that eight out of the nine LTMG-DEG specifically identified down regulated genes are truly down regulated ($p<1e-10$ and fold change<0.6), while the three MAST specifically identified DEGs are unchanged, in APEX1-KD vs control under the hypoxia condition. In addition, the qPCR analysis confirmed the up and down regulation for all the seven DEGs identified by both methods. These observations clearly suggest a better sensitivity and specificity of LTMG-DEG comparing to MAST.

We also examined the distribution of SR and AR peaks for the DEGs regulated by different TFs. Specifically, in APEX1-KD vs Control under the hypoxia condition, the LTMG-DEG identified that the DEGs regulated by STAT3 are with a higher proportion of SR expressions while the DEGs regulated by HIF1A are with a distinct AR peak with a smaller mean, suggesting a possible regulatory function loss of STAT3 and a lowered regulatory function of HIF1A in the APEX1-KD cells. Figure 5D-F shows the gene expression profile and inferred LTMG distribution of RARP4 regulated by STAT3 and SLC2A3 regulated HIF1A.

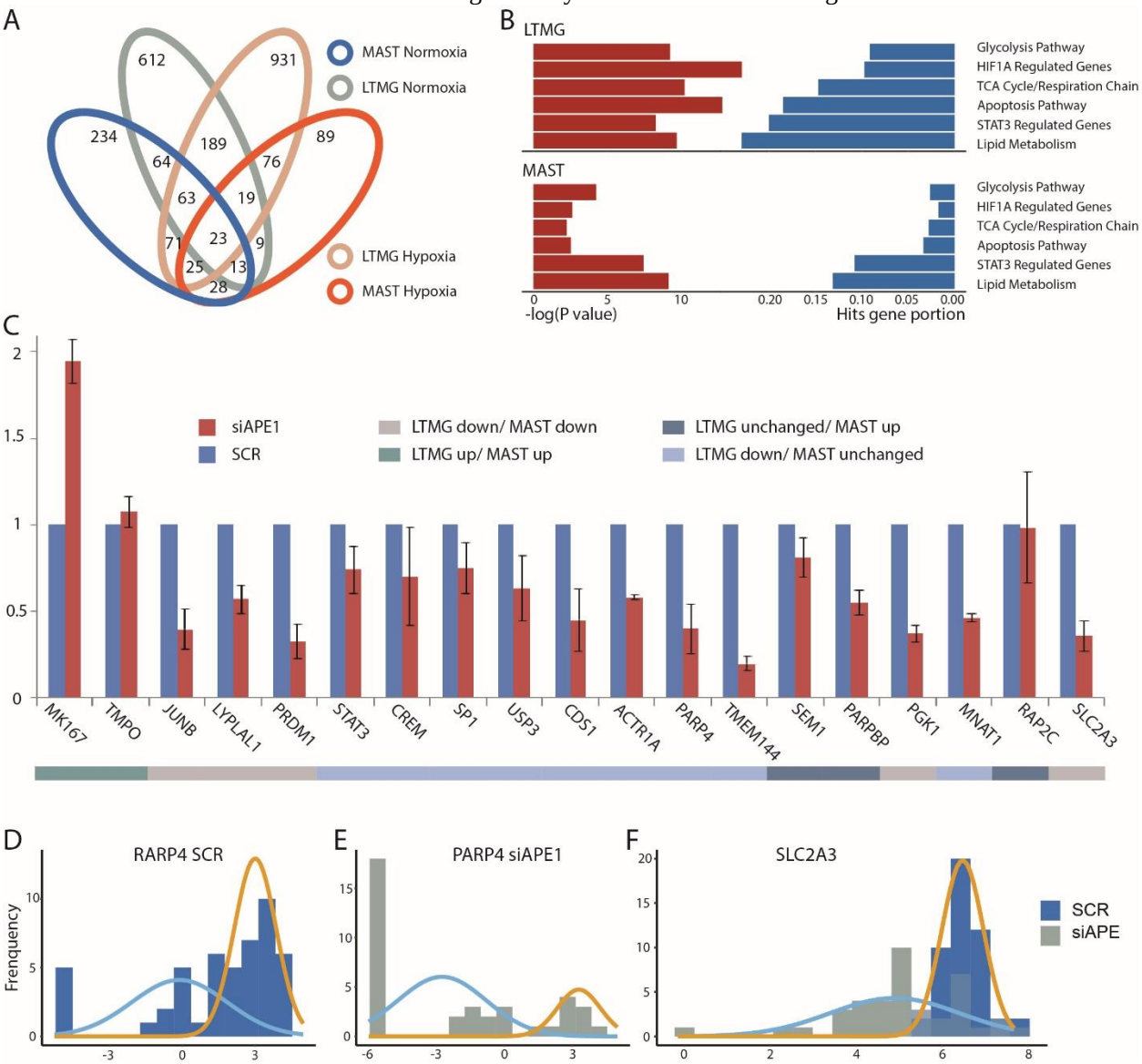

Figure 5. Experiment validation of LTMG-DEG. (A) Overlaps of down regulated genes in APEX1-KD vs Control in hypoxia and normoxia conditions, identified by LTMG-DEG and MAST. (B) Enrichment of the genes down regulated in APEX1-KD vs Control in key APEX1 related pathways, under the hypoxia condition. (C) qPCR results. (D-F) expression profile of RARP4 and SLC2A3 in APEX1-KD (siAPE) and Control (SCR) under the hypoxia condition. Gene expression level is quantified by log(RPKM) and represented on the x-axis.

LTMG-R – an R package for scRNA-seq data analysis within the LTMG scheme

With an inference of the heterogeneous TRS of each gene by our LTMG model, analysis of following problems for a scRNA-seq data, including (1) differentially expressed genes, (2) transcriptional bursts, (3) cell type specific gene regulations, (4) cell type inference, and (5) gene co-regulation module identification, can be formulated and computed by implementing LTMG model with other methods (Table 3). An R package of LTMG

model based single cell RNA-seq analysis tools was developed and released through github: https://github.com/zy26/coco-qubic. Specifically, the package includes the fitting of LTMG and LTMG-2LR models for single cell RNA-seq data, LTMG-DEG, LTMG based likelihood assessment, outlier gene and cell identification, identification of phenotypic feature associated TRS, cell type inference, and gene co-regulation module identification.

Table 3.

Question(s)	Biological definition	Formulation within the TRS framework	Computational Solution
Differential gene expression	Genes show differential expression	One gene is with one or several varied AR peaks in a group of cells comparing to others	Implement LTMG with a GLM or enrichment test (LTMG-DEG)
Transcription burst identification	High expression caused by a strong transcriptional activation	Expression of a gene that is significantly larger than its largest AR peak	Outlier identification based on LTMG
Cell type specific gene regulation	A TRS specific to a cell type	A gene with one or several SR/AR peaks that enrich a certain cell type	correlate the TRS inferred by LTMG with known cell types
Cell type inference	A group of cells form a possible cell type	Cells with a distinct number of cell-type specific TRSs	Implementation of LTMG with cell clustering algorithm
Gene co-regulation identification	Genes co-regulated by a TRS	A group of genes are with consistent SR or AR peaks in a certain subset of cells	Implement LTMG with bi-clustering

Discussion

LTMG model is developed for an accurate modeling of single gene's expression profile of scRNA-seq data, with considering the observed expression is determined by the transcriptional regulatory signals and mRNA metabolism in the current cell, and experimental resolution. To the best of our knowledge, LTMG is the first statistical model that considered the heterogeneity in transcription and metabolism rates of mRNA molecules in modeling scRNA-seq data. Our comprehensive model evaluation demonstrated that LTMG better handles the modality of high expressions, low expression caused by suppressed regulation and un-fully degraded mRNA, and have a much improved goodness of data fitting, comparing to other existing models in the public domain. Our experimental validation clearly suggested the LTMG-DEG is with a higher sensitivity and specificity than the tests based on other models.

A major challenge in modeling single-cell RNA-seq data is to simultaneously handle the multi-modality on the high expression side and the error term on the low expression side. Transcriptomic data of a bulk cell sample is the averaged sum of the mRNA abundance through all the cells. Varied gene expressions caused by heterogenous transcriptional regulatory signals in each cell are largely normalized in the total signal. Hence a unimodal distribution, such as the over dispersed Poisson or negative binomial distribution, can satisfactorily model a single gene's expression measured from the bulk samples a same population. On the contrary, the transcriptional regulatory signals through different cells, even for the ones of a same cell type, are always highly heterogenous due to dynamic states of the cells, which underlyingly determine the gene expression in each cell. Our mathematical derivation illustrated that the distributions such as zero-inflated Gaussian, Poisson with a beta-prior, and a mixture of two unimodal distribution are not enough to fit the expression profile of gene collected from cells with more than two TRSs. Previous studies have demonstrated that an over dispersed Poisson or negative binomial distribute can well fit the low expression part for unimodal cases. However, a dilemma is raised when extending this model to scRNA-seq data. The uncertain dispersion factor may cause a biased inference of the multimodal distribution on the high expression sides, i.e., parameter estimation of the AR peaks in Figure 1 depends on the over dispersion factor. Specifically, an AR peak can be absorbed by a large over dispersion factor that causes an under estimated number of AR peaks. On the other hand, our LTMG model

considers the errors of the low expressions cannot be accurately estimated, which are modeled as censored data unsuccessfully detected by the experiment resolution, and are possibly from a SR peak and a number of AR peaks. The model provides a solution to simultaneously handle the uncertain error of low expressions and inference of the multimodality of the high expressions. We also compared LTMG with ZIMG in our method evaluation. It is worth to note that the ZIMG model infers AR peaks based all non-zero expressions including both true AR expressions and low expression from SR peak (caused by un-fully degraded mRNA). Hence ZIMG overly fits the data and tends to estimate a higher number of AR peaks. Our comparison demonstrated a better fitting of LTMG comparing to ZIMG in more than 90% analyzed data sets, especially for the ones with large sample size, further confirms the rationale of our LTMG assumption.

We have experimentally validated the LTMG-DEG has better sensitivity and specificity comparing to MAST, due to the gene expression profile is more accurately fitted by the LTMG model. In addition, we note that LTMG-DEG utilizes a more rigorous GLM formulation for differential expression test than MAST, due to it models SR expression as a true Gaussian comparing to a Gaussian truncated by the experimental resolution utilized by MAST.

It is noteworthy that the LTMG model is designed for the scRNA-seq with a large sequencing amount for each cell, i.e. distribution of the normalized gene expression profile is not affected by variations in total sequencing reads. However, such an assumption does not fit the drop-seq based scRNA-seq data since there is always a big span of total number of reads through the cells and the total number of reads for each is far less than level of a saturated resolution. SC2P introduced a concept to model scRNA-seq with a consideration of cell wise sequencing resolution. A possible future direction is to implement the cell wise factor into the LTMG model for drop-seq data such as 10x genomics data.

Another key challenge in scRNA-seq analysis to develop a biologically meaningful dimension reduction formulation for gene and cell cluster, which encounters (1) select only genes which are cell type / biological function deterministic, (*ii*) identify the rational low rank patterns that encodes the true characteristics of a cell type. LTMG naturally links the TRS of different genes to a bi-cluster structure (see more details in the LTMG R package). It is noteworthy that the bi-cluster formulated in this way corresponds to a group of gene, each of which is regulated by one specific AR signal, in a certain subset of cells, which possibly indicate a gene co-regulation module, i.e. a group of genes co-regulated by a common TRS to multiple cells. ScRNA-Seq data is an ideal data type for gene regulation mechanism study due to each gene's expression in single cell is a manifestation of its currently effective TRS. We anticipate a further implementation of LTMG model with a bi-clustering approach can shed new light to single cell transcriptomics data based gene regulation study.

Data and Methods

Data Collection

Datasets were collected based on three criteria: 1) human or mouse data, 2) with a sample size between 100 to 6400, 3) generated by using SMART-seq or SMART-seq2 protocol., by which 20 datasets were retrieved from GEO database. To evaluate the performance of the LTMG model in the data sets of different level of sample complexity, we extracted sub datasets from the collected data, and classified them into three subgroups by the homogenous levels of cell samples, namely (1) Pure Condition (PC), defined by the data of one unique cell line or cell type under a specific experimental condition; (2) Cell Cluster (CC), for the data of computationally clustered cells; and (3) Complete Data (CD), as the data of a mixed cell population such as cells from one cancer tumor tissue or a complete data. In total, 51 PC, 49 CC, and 78 CD data sets were identified from the 20 large data sets.

Fitting of LTMG model by an EM algorithm

An EM algorithm to estimate the parameters $\Theta = (a_i, u_i, \sigma_i | \, i = 1 \ldots K)$ of an LTMG model is developed by using the following Q function, with introducing the latent variables y_i and Z_j, and prespecified K and Z^X_{cut}:

$$Q(\Theta, \Theta^{t-1}) = \sum_{i=1}^{K} \sum_{j=1}^{M} log(a_i p_i(x_j | u_i, \sigma_i) p(y_j = i | x_j, \Theta^{t-1}) + \sum_{j=M+1}^{N} \int_{-\infty}^{Z^X_{cut}} \sum_{i=1}^{K} log(a_i p_i(Z_j | u_i, \sigma_i)) p(y_j = i | Z_j, \Theta^{t-1}) p(Z_j | x_j, \Theta^{t-1}) dZ_j$$

, where Θ represents the to-be-estimated parameters, Z_j is the latent variable reflecting the real expression level of x_j if the measured expression level is smaller than Z_{cut}^X, $y_j = 1 \ldots K$ is a latent variable indicating that x_j is the from the y_jth Gaussian peak, and t is the current iteration step.

With the above Q function, we have derived an EM algorithm to estimate the parameters Θ that maximizes the likelihood function, with the parameters iteratively updated in the following Maximization (M) step:

$$a_i^t = \frac{1}{N}\left(\sum_{j=1}^{M} P\left(i \middle| x_j, \Theta^{t-1}\right) + \sum_{j=M+1}^{N} P(i|Z_j, Z_{cut}, \Theta^{t-1})\right)$$

$$u_i^t = \frac{\sum_{j=1}^{M} x_j P\left(i \middle| x_j, \Theta^{t-1}\right) + \sum_{j=M+1}^{N}\left(u_i^{t-1} - \sigma_i^{t-1} H\left(\frac{Z_{cut} - u_i^{t-1}}{\sigma_i}\right)\right) P(i|Z_j, Z_{cut}, \Theta^{t-1})}{\sum_{j=1}^{M} P(i|x_j, \Theta^{t-1}) + \sum_{j=M+1}^{N} P(i|Z_j, Z_{cut}, \Theta^{t-1})}$$

$$\sigma_i^{t^2} = \frac{\sum_{j=1}^{M} P\left(i \middle| x_j, \Theta^{t-1}\right)(x_j - u_i^{t-1})^2 + \sigma_i^{t-1^2} \sum_{j=M+1}^{N}\left(1 - \frac{Z_{cut} - u_i^{t-1}}{\sigma_i} * H\left(\frac{Z_{cut} - u_i^{t-1}}{\sigma_i}\right)\right) * P(i|Z_j, Z_{cut}, \Theta^{t-1})}{\sum_{j=1}^{M} P(i|x_j, \Theta^{t-1}) + \sum_{j=M+1}^{N} P(i|Z_j, Z_{cut}, \Theta^{t-1})}$$

, where $P\left(i \middle| Z_j, Z_{cut}, \Theta^{t-1}\right) = \frac{P(-\infty < Z_j < Z_{cut} | u_i^{t-1}, \sigma_i^{t-1})}{\sum_{i=1}^{K} P(-\infty < Z_j < Z_{cut} | u_i^{t-1}, \sigma_i^{t-1})}$, $H(x) = \frac{\phi(x)}{\Phi(x)}$, $\phi(x)$ and $\Phi(x)$ are the pdf and cdf of standard normal distributions.

LTMG-2LR distribution

AN LTMG-2LR distribution is defined by an LTMG distribution with two peaks, and the mean of one peak is smaller than Z_{cut} while the mean of the other peak is larger than Z_{cut}, i.e. $\Theta = \{a_i, u_i, \sigma_i | i = 1, 2\}$, $u_1 < Z_{cut}$ and $u_1 > Z_{cut}$.

BIC based peak number selection

Parameters Θ can be estimated by iteratively running the estimation (E) and maximization (M) steps. In this study, Z_{cut}^X is set for X as the logarithm of the minimal non-zero value in the gene's expression profile, i.e. $Z_{cut}^X = \log(\min(X_i | X_i > 0))$. The EM algorithm is executed for K = 1, …, 5 to fit the expression profile of each gene, and the K gives the best of fitting is selected according to the Bayesian Information Criterion (BIC):

$$BIC = -2\ln(\Theta^*) + 3K\ln(N)$$

, where N is the number of cell samples.

Model application and fitting comparison

We applied Zero-inflated mixed Gaussian (ZIMG), Left Truncated Mixed Gaussian (LTMG), MAST and Beta Poisson (BPSC) on each classified dataset. We use MAST with default parameters. After filtering, the remaining positive values were fitted with Gaussian distribution. As for BPSC, to achieve a reliable estimation, we only fitted genes with over 25 positive counts, such that, 88 datasets were fitted with BPSC. Kolmogorov Statistic (KS) is used to measure gene-wise goodness of fitting. The KS score is normalized by the proportion of zero observations for the ZIMG, MAST and BPSC models, due to their zero inflation assumption. We excluded the genes failed to be fitted by any one of the four models from the downstream analysis. Fitting comparison of each datasets is based on the mean and variance of gene-wise KS values.

GF score, the goodness fitting measurement on dataset level

To measure the goodness fitting on the dataset level, we defined a GF score by

$$GF_{score} = \frac{1}{2}\left(\overline{KS} + \sigma(KS)\right)$$

, where $\overline{KS}$ is the mean value of KS scores from a single dataset and $\sigma(KS)$ is the standard deviation of the KS scores. The less the GF score is, the greater the fittings are on the dataset level, as the fitting performs comprehensively well on both accuracy (lower $\overline{KS}$ value) and stability (lower $\sigma(KS)$). Noted that some

complete datasets were divided and classified into many other datasets, GF score of these data was accordingly computed by using the mean value of GF scores of all the sub datasets.

MRNA kinetics and the mouse fibroblast datasets

Kinetics of mRNA in mouse fibroblast cells are collected from Björn Schwanhäusser et al's work. We manually curated three datasets (GSE99235, GSE98816 and GSE87038) of mouse fibroblast cells to test the association between the LTMG fitted parameter and the experimentally measured mRNA half-life. Cells in each data set were first clustered by using Seurat with default parameters. Sub population of fibroblast cells were identified by expression of collagen and other fibroblast specifically expressed genes [24]. We identified 369 fibroblast cells in GSE87038 (cluster 2 and 3), 397 fibroblast cells in GSE99235 (cluster 1) and 1100 fibroblast-like cells in GSE98816 (cluster 1 and 2) data. LTMG model were applied on the data of these cells. Fitted parameters were correlated with experimentally measured mRNA half-life.

Correlation analysis of LTMG with mRNA metabolism kinetics

In our hypothesis, AR peak with lower mean has a higher impact to the measurement of the uncensored region in SR peak part (greater type II error). In order to derive a general trend between the region proportion and mRNA half-life, we measure the correlation between the mRNA half-life and proportion of uncensored expression in the SR peak, condition to the mean of AR peaks, i.e. level of type II error. Specifically, we first split the genes into groups based on mean of their first AR peak. Here we select a slice size of 100 due to a consideration of the following tradeoffs: (1) for each sliced group, a low within group variation is needed to ensure the genes within each group have a similar level of AR peak type II error; and (2) a relatively bigger slice size is needed to guarantee the significance of correlations. For each dataset, we order every gene from low to high by the mean of the first AR peak, slice them by every 100 genes, and compute the correlation between the mRNA half-life and proportion of uncensored expressions in each group. Students' T test for the moment of Pearson Correlation Coefficient is used to assess the significance of correlations and $p<0.1$ is used a significance cutoff.

E score and analysis of cell type specifically expressed genes

We designate cells into SR or AR peak by probability mapping. Compared with all peaks, cells are designated to a peak which they have the greatest probability to appear. With prior knowledge of cell type information, we could calculate the proportion of each cell type within a single peak. E score is then defined as the exponential value of the proportion of a specific type of cell within that peak.

T-SNE visualization *of the head and neck cancer*

We clustered GSE103322 datasets by using the Rtsne package with 30 complexity and 20000 max iterations. We only used the markers genes provided by the original paper for cell clustering. The t-SNE analysis is only for data visualization. Cell type annotated in the original paper was used to label the cell types in Figure 4.

LTMG based differential gene expression analysis

Denote $X_j = \{x_i^j, i = 1 \dots N_j\}, j = 1 \dots J$, as gene X's expression profile in the N_j cells of the jth conditions, where J is the number of to-be-tested conditions. Since the number of Gaussian peaks in the LTMG model is determined by BIC, a rigorous statistical test of the association between a TRS and a condition can be made only if the total number of TRSs, i.e. the number of Gaussian peaks, is pre-specified [25].

Noting that more than 80% of the genes in the PC and CC groups are fitted with one or two peaks, and more than 90% of the genes fitted with two peaks are actually fitted with an LTMG_2LR model, we developed an LTMG_2LR based test for the genes that are fitted with no more than two peaks in all the to-be-tested conditions. For $X_j, j = 1 \dots J$, we first fit an LTMG_2LR for each X_j with assuming a same parameter set (u_0^X, σ_0^X) of the SR peak through all the conditions as shown below:

$$\begin{cases} \mathrm{X}_1 \sim LTMG_2LR(a_1^X, u_0^X, u_1^X, \sigma_0^X, \sigma_1^X) \\ \mathrm{X}_2 \sim LTMG_2LR(a_2^X, u_0^X, u_2^X, \sigma_0^X, \sigma_2^X) \\ \mathrm{X}_3 \sim LTMG_2LR(a_3^X, u_0^X, u_3^X, \sigma_0^X, \sigma_3^X) \\ \ldots \end{cases}$$

The rationality of this assumption is that one gene may share a same SR state and with relatively similar degradation rates through different conditions. Then differences in $a_1^X, \ldots, a_J^X$ and $u_1^X, \ldots, u_J^X$ can be rigorously tested by implementing a GLM model with a random sampling process as detailed below. With $\Theta^{\mathrm{X}} = （a_j^X, u_j^X, \sigma_j^X, u_0^X, \sigma_0^X | j = 1 \ldots J）$ estimated, the probability that x_i^j belongs to a SR (or AR) peak can be assessed, denoted as $\mathrm{p}(\mathrm{x}_i^j \in SR) = 1 - \mathrm{p}(\mathrm{x}_i^j \in AR)$. A sampling process can be made by randomly assigning x_i^j to the SR (or AR) state of condition j with probability $\mathrm{p}(\mathrm{x}_i^j \in SR)$ (or $\mathrm{p}(\mathrm{x}_i^j \in AR)$), by which $a_1^X, \ldots, a_J^X$ can be tested by using a logit linking function to link the frequency of x_i^j belong to the SR (or AR) state under each condition, with the design matrix of the conditions; and $u_1^X, \ldots, u_J^X$ can be tested by using a linear linking function to link the mean of x_i^j belong to the AR state under each condition, with the design matrix. With applying the random sampling process N times, p value of each test is estimated by the median of the identified p values, and the confidence intervals of each p value can be estimated. The advantages of this process include (1) rigorousness of the GLM form, (2) high sensitivity for the changes in frequency or mean expression level of the AR state, and (3) the testing rigorousness is not affected by the dilemma of a mixture distribution, due to u_0^X and σ_0^X are fixed for all conditions.

For a gene that is fitted with more than two AR peaks in at least one condition, to avoid the dilemma of a mixture distribution, we applied the following enrichment test by pre-specifying all the TRS of the gene: (1) fit an LTMG model by using the pooled data of all conditions, i.e. $\mathrm{X} \sim LTMG(a_i^X, u_i^X, \sigma_i^X \mid i = 1 \ldots K)$, $\mathrm{X} = \{\mathrm{x}_i^j, \mathrm{i} = 1 \ldots \mathrm{N_j}, \mathrm{i} = 1 \ldots \mathrm{J}\}$, (2) compute the likelihood that x_i^j belongs to peak $i = 1 \ldots K$ and assign x_i^j to the peak with the maximal likelihood, (3) compute if the samples of each condition $j = 1 \ldots J$ are enriched to a peak $i = 1 \ldots K$ via a hypergeometric test.

The difference of the two testing schemes is that the former one assumes a gene is only with one AR TRS in each condition, but can be varied in proportion, mean, or variance through the conditions, and the test is for the proportion and mean, while the later one first fixed the distribution of each TRS by a fitting of the pooled data through all conditions, and test if one condition is specifically associated with one TRS. In sum, the LTMG-DEG is formed by two steps: (1) testing if a gene can be well fitted with LTMG_2LR model through all conditions, (2) if yes, conducting the GLM and random sampling based test for $a_1^X, \ldots, a_J^X$ and $u_1^X, \ldots, u_J^X$; else, fitting an LTMG model with the pooled data, and applying the enrichment test of the TRSs.

ACKNOWLEDGMENT

This work is supported by NIH NIGMS, 1R01GM131399-01 and Showalter Young Investigator Award, Indiana CTSI. CZ would like to specifically thank Dr. Yunlong Liu, Dr. Kun Huang, Dr. Jun Wan and Dr. Kelley Mark at IU school of Medicine for their support in the single cell data generation and analysis of this work.

REFERENCES

1 Puram, S. V. *et al.* Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611-1624. e1624 (2017).

2 Azizi, E. *et al.* Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293-1308. e1236 (2018).

3 Li, L. *et al.* Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858-873. e854 (2017).

4 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).

5 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* **9**, 171 (2014).

6 Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).

7 Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature methods* **11**, 740 (2014).

8 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology* **16**, 278 (2015).

9 Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**, 2128-2135 (2016).

10 Wu, Z., Zhang, Y., Stitzel, M. L. & Wu, H. Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics* **1**, 9 (2018).

11 Johnson, K. A. & Goody, R. S. The original Michaelis constant: translation of the 1913 Michaelis–Menten paper. *Biochemistry* **50**, 8264-8269 (2011).

12 Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29-59 (2006).

13 Voss, T. C. & Hager, G. L. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics* **15**, 69-81 (2014).

14 Wang, J., Tsang, W. W. & Marsaglia, G. Evaluating Kolmogorov's distribution. *Journal of Statistical Software* **8** (2003).

15 Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337 (2011).

16 He, L. *et al.* Single-cell RNA sequencing of mouse brain and lung vascular and vessel-associated cell types. *Scientific data* **5**, 180160 (2018).

17 Dong, J. *et al.* Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome biology* **19**, 31 (2018).

18 Barry, M. & Bleackley, R. C. Cytotoxic T lymphocytes: all roads lead to death. *Nature Reviews Immunology* **2**, 401 (2002).

19 Guo, Y., Chen, J., Zhao, T. & Fan, Z. Granzyme K degrades the redox/DNA repair enzyme Ape1 to trigger oxidative stress of target cells leading to cytotoxicity. *Molecular immunology* **45**, 2225-2235 (2008).

20 Wherry, E. J. T cell exhaustion. *Nature immunology* **12**, 492 (2011).

21 R Kelley, M., M Georgiadis, M. & L Fishel, M. APE1/Ref-1role in redox signaling: translational applications of targeting the redox function of the DNA repair/redox protein APE1/Ref-1. *Current molecular pharmacology* **5**, 36-53 (2012).

22 Shah, F. *et al.* APE1/Ref-1 knockdown in pancreatic ductal adenocarcinoma–characterizing gene expression changes and identifying novel pathways using single-cell RNA sequencing. *Molecular oncology* **11**, 1711-1732 (2017).
23 Cardoso, A. A. *et al.* APE1/Ref-1 regulates STAT3 transcriptional activity and APE1/Ref-1–STAT3 dual-targeting effectively inhibits pancreatic cancer cell survival. *PloS one* **7**, e47462 (2012).
24 Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6**, e26476 (2017).
25 Claeskens, G. & Hjort, N. L. Model selection and model averaging. *Cambridge Books* (2008).