NonLoss: a novel analytical method for differential biological module identification from single-cell transcriptome

Hui Zhao[1,2,3], Ying Guo[1,2], Yanan Ma[1,2], Yunping Chen[1,2], Haiming Sun[1,2], Donglin Sun[1,2], Nan Wu[1,2], Yan Jin[1,2]

[1]Key Laboratory of Preservation of Human Genetic Resources and Disease Control in China (Harbin Medical University), Ministry of Education, Harbin, China; [2]Laboratory of Medical Genetics, Harbin Medical University, Harbin, China; [3]Department of Hematology, The First Affiliated Hospital, Harbin Medical University, Harbin, China

Contributions: (I) Conception and design: H Zhao, N Wu; (II) Administrative support: Y Jin; (III) Provision of study materials or patients: H Zhao, Y Guo, Y Ma; (IV) Collection and assembly of data: H Zhao, Y Chen, H Sun, D Sun; (V) Data analysis and interpretation: H Zhao, N Wu, Y Jin; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Nan Wu; Yan Jin. Laboratory of Medical Genetics, Harbin Medical University, 157 Baojian Road, Harbin 150081, China. Email: wunan@ems.hrbmu.edu.cn; jinyan@ems.hrbmu.edu.cn.

Background: The identification of disease-related biological modules plays an important role in our understanding of the process of diseases. Although single-cell RNA sequencing (scRNA-seq) provides high-resolution transcriptome data that can potentially characterize subtle gene expression changes within cells, the susceptibility of the gene expression information to the influence of individual genes also makes it difficult to distinguish the biological module.

Methods: To quantify gene expression information for biological function modules, we adopted the method based on Shannon's entropy and Spearman rank correlation analysis. The ingenious combination of these two methods enables the variation analysis of the former and the consistency analysis of the latter to make a more robust biological function analysis tool.

Results: We developed a computational analytical method and desktop application called NonLoss to analyze scRNA-seq data more robustly and to extract real biological differences between cell populations. The method derives its power by handling expression level data from all genes annotated to a specific function module, both for dimensionality reduction and reliability of function identification, avoiding random disturbance of individual genes. NonLoss can in principle be used to assess changes of function modules and identify vital functions simultaneously. Furthermore, specific genes contributing to important functions, even those with subtle expression changes, can be identified. The results demonstrated that NonLoss yields biologically significant insights into 3 different applications.

Conclusions: NonLoss was developed with a user-friendly graphical user interface, and it could identify the module of biologically relevant expression changes at a single-cell resolution.

Keywords: Single-cell RNA sequencing (scRNA-seq); biological module; blood; colorectal cancer; python package

Submitted Nov 05, 2021. Accepted for publication Dec 22, 2021.
doi: 10.21037/atm-21-6401
View this article at: https://dx.doi.org/10.21037/atm-21-6401

Introduction

Cells are the smallest unit of individual life, but they are still exquisite at the molecular level. Biological processes within cells are dynamic and embodied at gene transcriptional level. However, the bulk of RNA-seq largely focuses on quantifying the gene expression information across a heterogeneous population of cells (1). Single-cell RNA sequencing (scRNA-seq) is a powerful, high-resolution tool used for the study of cellular heterogeneity at a transcriptome level, in individual cells (2). It has been used to analyze embryonic development, cancer heterogeneity, and even novel cell types (3,4).

According to the complexity of data analysis of cellular dynamics, biological variability of individual cells, the presence of technical limitations, and library size or composition bias, discerning the real biological differences (such as changes in biological module regulation or gene expression levels) between cells remains challenging. One of the most commonly performed tasks for RNA-seq data is gene differential expression (DE) analysis (5). However, due to the uneven gene coverage of scRNA-seq across different cells or experiments, even those well-established tools for such analysis produce bias inevitably when calling the DE genes (6-8), resulting in misleading conclusions of biological function analyses. To make scRNA-seq analysis protocols more robust, tremendous efforts have been made such as the spike-in RNA approach for scaling normalization and the unique molecular identifiers (UMI) method to avoid amplification biases from experimental angles (9). To some extent, these methods are beneficial, yet limited. Ultimately unstable factors still lead to DE genes bias when applied to scRNA-seq data. Conclusively, we should be aware when DE analysis was applied to the exploration of biological mechanisms and functions. From this perspective, gene expression data has been analyzed on a gene-by-gene basis by DE analysis, without regard for the overall impact on biological functions. Even more detrimentally, genes with lower expression ratios [fold change (FC) value <2, the major part of genes] are routinely excluded from downstream gene enrichment analyses (10,11). Furthermore, Various other approaches like D3E, MAST and SigEMD have been developed for the DE analysis (6,12,13). However, these tools try to deal with either the gene dropouts or multimodality. For the subtle DE genes as well as weak expressed genes were ignored or discounted, and the accumulation of information may be even more important.

Thus, due to the potential liability of gene expression levels, the strategy of comparing the same gene across different cells for scRNA-seq data could be meaningless. Alternatively, scRNA-seq provides more detailed information to build a complete map of the transcriptome of a single-cell, exposing the cell state diversity and perturbations in vivid detail (14,15). To increase stability of functional discovery, aggregating related functional genes and measuring them collectively could be a viable option. We propose a function-oriented, non-statistical approach (referred to as NonLoss), to identify differential biological modules (DBMs) between cells or conditions. We categorized genes as well as gene expression data into biological function modules (BFMs) (also known as gene functional annotation), and then estimated expression differences based on the functional units rather than the individual genes. The NonLoss strategy successfully curtails the arbitrary effects of FC cutoffs of DE methods and aggregates a composite of weak evidence to identify functional significance, enhancing the power of transcriptomics at the single-cell resolution for understanding multiple biological processes. To facilitate the use of NonLoss, we have developed a software package that is freely available from https://github.com/X1angyang/Nonloss-V1 upon request. The software is available as a desktop application with a graphical user interface and is programmed in Python. A detailed example of input and output data format of NonLoss is available in the help documents.

We present the following article in accordance with the MDAR reporting checklist (available at https://dx.doi.org/10.21037/atm-21-6401).

Methods

Biological function annotation

The scRNA-seq technique provides information in a high dimensional gene expression space. When comparing cells in a high dimensional gene expression space, distances between cells become more homogenous and a small significant difference can be easily overwhelmed by large volumes of expression data, making it difficult to distinguish differences between cell populations. Our main objective was to explore the changing trends of biological functions rather than several DE genes. Genes in the same biological module tend to exhibit strong corresponding changes at a transcriptional level. Gene Ontology (GO) provides structured, controlled vocabularies and classifications of categories (16,17). Exploring GO annotations for insights into the potential experimental meanings has become a widespread practice. Therefore, we chose GO as the biological function module (BFM) to demonstrate our analytic approach. However, our work can also be applied to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis and other user-defined biological function sets. Since computing work causes instability when applied to BFM with a small number of genes, we set the lower gene number limit to 10. Meanwhile, the BMF variation stabilized as the gene number of BMF increased. The upper gene number limit was set to 200. Of note, users can freely adjust the upper and lower limits depending on their experimental needs.

Transcriptome reference construction

Previous studies using conventional bulk RNA-seq have handled a relatively small number of samples with explicit sectioning. However, the nature of scRNA-seq is that it generates from thousands of samples in a single experiment with high-noise attributes (18). Additionally, the goal of the study was to explore molecular dynamic changes or cellular specific traits of individual cells. So, we could neither perform analysis by simply taking average comparisons between conditions nor compare 2 cells in arbitrary ways. To deal with this problem, inter-group reference (IGR) of a cell population was proposed to provide a benchmark for assessing differences among cells. The IGR was constructed by picking the median of gene expression level for each group of cell population (for example IGR of case group or IGR of the control group). Then, each cell from one group was compared with an IGR of another group by our NonLoss method. The relative BFM states of each cell were presented thoroughly.

Information measurement model based on Shannon's entropy

To quantify gene expression information for BFMs, we adopted the method based on Shannon's entropy. Although entropy has previously been used to identify DE genes for gene expression profiles (19), we were the first to apply it to quantify BFMs with more complete gene expression information. To make the description easier to follow, we measured the gene information related to specific GO terms using Shannon's Entropy. The CRC datasets was used for method building and process testing (20). Considering that e_i^j denotes the gene expression vector of the GO term $i,\left(i \in List^{GO}\right)$ of cell j, and $List_i^{Gene}$ was gene list annotated in the GO term i, then the sum gene expression value of e_i^j was calculated as $S_i^j = \sum List_i^{Gene} e_i^j$. Therefore, the information entropy of GO term i of cell j can be measured as

$$H_i^j = -\sum_{List} Gene \frac{e_{i,k}^j}{j} \log_2\left(\frac{e_{i,k}^j}{j}\right), k \in List_i^{Gene} \qquad [1]$$

In information theory, entropy is a measure of the uncertainty associated with a random variable (21). In this context, Shannon's entropy was used to quantify the expected value of the information contained in a BFM or GO term. Although entropy is often used as a characterization of the information content of a data source, this information content depends crucially on the probabilistic model. We assumed that a major part of gene expression level will not fluctuate too much relative to its normal state. The probability distribution of gene expression was relatively stable for pairwise comparisons. So, the Shannon's entropy value can be fixed to some extent, and then compared. However, the entropy functions follow the property that H_i is an increasing function of gene number N^i of GO term i, (*Figure 1A*). To overcome this shortcoming, the entropy model should be further normalized to account for the unequal number of GO terms, thus the standard entropy difference (SED) can be defined as

$$\mathrm{SED}_\mathrm{i}^\mathrm{j} = \frac{H_i^{cell_j} - H_i^{IGR}}{\log_2(N^i)} \qquad [2]$$

After normalization, this problem can be solved reasonably using this formula (*Figure 1B*). The higher SED_i^j the greater GO difference was represented compared with IGR. We used Shannon's entropy to measure the information difference with more stability and less susceptibility to the influence of stochastic events (22), thereby reducing the impact of non-functional related DEGs in BFMs. We did not expect a single gene to become the highlight; instead, we aimed to discover a set of genes that affected biological functions. This method was relatively insensitive to outliers. However, the Shannon entropy model did not consider the internal expression order in a BFM. For example, if 2 genes A and B were annotated to a GO term k and the gene expression vector of 2 cells m and n was $e_k^m = \left(e_k^{m,A} = 0.2, e_k^{m,B} = 0.8\right), e_k^n = \left(e_k^{n,A} = 0.8, e_k^{n,B} = 0.2\right)$, respectively. Despite the great differences between the 2 at a transcriptional level, the value obtained from the formulas would be $SED_k = 0$ resulting in an incorrect conclusion. To overcome this shortcoming, we introduced consistency analysis which is defined in the subsequent sub-section.

Consistency analysis

Entropy has a problem with measuring the pertinence of the variable elements within a calculation unit. To solve this problem, we introduced the Spearman rank correlation

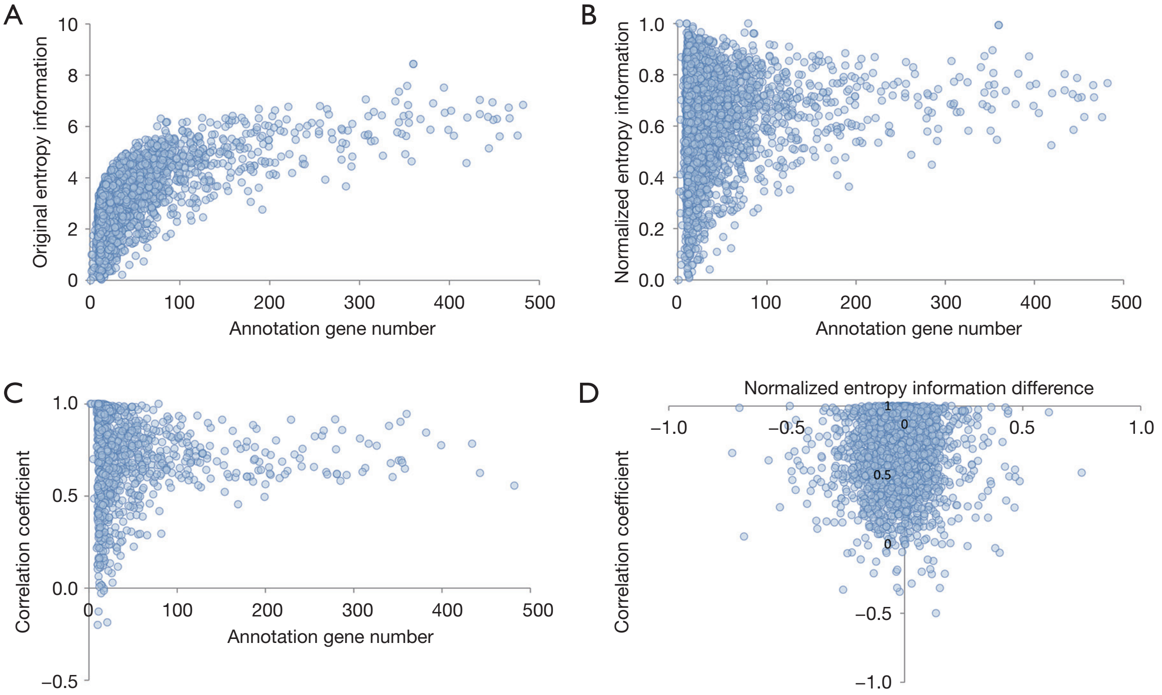

Figure 1 Scatter diagrams used to show the characteristics of entropy information and correlation coefficients of tested GO terms. (A) Scatter plot of annotation gene number (x-axis) and original entropy information (y-axis), entropy information increased with the increase of annotation gene number. (B) Scatter plot of annotation gene number (x-axis) and normalized entropy information (y-axis). (C) Scatter plot of annotation gene number (x-axis) and correlation coefficient (y-axis). The correlation coefficients were more volatile with smaller annotation gene numbers and became stable with the increase of annotation gene number. (D) Scatter plot of normalized entropy information difference (x-axis) and correlation coefficient. The IScores were computed by the Euclidean distance between coordinated point values of GO terms and coordinates [0, 1]. GO, Gene Ontology; IScore, integrated scores.

analysis, a non-parametric similarity measure which is robust against outliers (23). This compensated for the disadvantage of the original entropy method to measure the consistency of gene expression in BFMs.

Correlation simply measures the relationship of gene expression consistency between 2 samples for a specific BFM. This relationship, which was expressed by what is known as the correlation coefficient, is represented by a value within the range of (−1.0, +1.0). To achieve significant differences, genes in specific BFM between 2 samples should not be highly correlated. When the correlation increased, the diversification difference decreased and vice versa. A correlation coefficient of +1.00 indicates that the expression of genes in a specific BFM showed a perfect proportion between 2 cells. This module always gives a very weak difference between samples. A correlation coefficient of 0 indicates that expressions of genes in a specific BFM are completely random. This may be attributed to biological molecules function disorders. A correlation coefficient of −1.00 indicates that expression of genes in a specific BFM show the opposite direction between 2 samples. This may be an interesting result depending on the annotation gene number. The correlation coefficients when annotation gene numbers were small were volatile and stabilized as the annotation gene numbers increased. The probability of this case was very small if an appropriate threshold was selected (*Figure 1C*).

Integrated score with IScore and difference of IScore

As mentioned, both the entropy information method and the consistency analysis method can explain part of variations for biological function. Through analysis, it was found that they complement each other very well. In this section, we proposed a comprehensive method to measure the degree of GO difference by combining the entropy method and consistency method. We introduced Euclidean distance to measure the degree of BFM variation, which was defined as follows:

$$IScore_j = \sqrt{SED_i^2 + (R_i - 1)^2} \quad [3]$$

Therefore, the integrated scores (IScore) of BFM were defined as the Euclidean distance between (SED_i,R_i–1) nd (0,1) (*Figure 1D*). Then, the BFM variations of each cell were quantified by comparing the IGR using IScores. The IScore was primarily used to compare biological functional changes between 2 cell populations. Thus, the difference of IScore (DIS) was defined by the taking difference of two population averages from each BFM.

$$DIS = average(IScore_a) - average(IScore_b) \quad [4]$$

Statistical analysis

All statistical tests were performed using Python3.7 (scipy 1.4.1). Correlation coefficient were calculated by spearman correlation. Difference between two groups were tested using unpaired *t*-test and FDR P value <0.05 were considered statistically significant.

Data sets description

The first data set [Gene Expression Omnibus (GEO) accession number: GSE81861] contained 375 cells collected from colorectal cancer (CRC) tissue and 215 nearby normal mucosa qualified expression profiles (20). A total of 272 tumor and 160 normal mucosa epithelial cells were identified respectively by ring clustering algorithm (RCA) (20). We explored biological feature differences in these 2 cell populations. For ease of description, CRC-data was defined as an abbreviation for the first data set. A second data set was from human non-stimulated and cytokine-activated mucosal-associated invariant T cells (MAIT) (6). The MAIT dataset was defined as an abbreviation for the data set. The sample came from human non-stimulated and cytokine-activated [interleukin (IL)-12, IL-15, and IL-18 treated] MAIT cells which were from peripheral blood mononuclear cells. The $CD8^+$ MAIT cells were sorted. We obtained data from 47 stimulated and 49 non-stimulated MAIT profiles and used them for further functional analyses. The primary human skeletal muscle myoblasts (HSMM) data sets (GEO accession number: GSE52529) containing both scRNA-seq and bulk RNA-seq data were used as the third data set for validating our method and extending the application with a bulk RNA-seq data (24). In each analyzed scRNA-seq dataset, genes that were never expressed were filtered out. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Results

We aimed to acquire insight on the gain or loss of biological functions of different cell types and provide candidate targets for further investigation. To examine the newly developed DBMs identification pipeline, 3 published scRNA-Seq datasets were applied (see Methods section). The first CRC data was used to perform systematic and comprehensive analyses for purposes of method validation and tutorial construction. NonLoss takes gene annotations and gathers all gene expression information for consideration. As its working principle was different, it could not be directly compared with other traditional DE methods. Then, an alternative strategy was used to compare the findings of NonLoss and MAST methods (6). The second data set (MAIT) was used for the sake of consistency and reproducibility in comparison to traditional DE method. Nevertheless, the scRNA-seq data was expected to uncover more sensitive biological features relative to bulk RNA-seq. We expected to find overlap DBMs between the 2 sequencing schemes. For this reason, the third data set [human skeletal muscle myoblast (HSMM)] containing both scRNA-seq and bulk RNA-seq data were used for validation of NonLoss, as well as extending the application to bulk RNA-seq data. We envisioned that NonLoss could be useful for a range of applications. The detailed information for these 3 data sets is presented in the Methods section.

Validation of NonLoss through application to CRC data set

Differential KEGG pathways identification

As KEGG is a comprehensive and reliable knowledge base for assisting biological interpretations of large-scale molecular datasets, to explore significant biological functional modules and interactions between tumor and normal mucosa epithelial cells of CRC-data, we first carried out NonLoss analysis based on KEGG pathway modules.

In this case, parameter default values were used to calculate the IScore of each cell. The states of specific biological functions for each cell were quantified by IScore. Therefore, these IScore values served to reveal the heterogeneity of biological functions of the cell

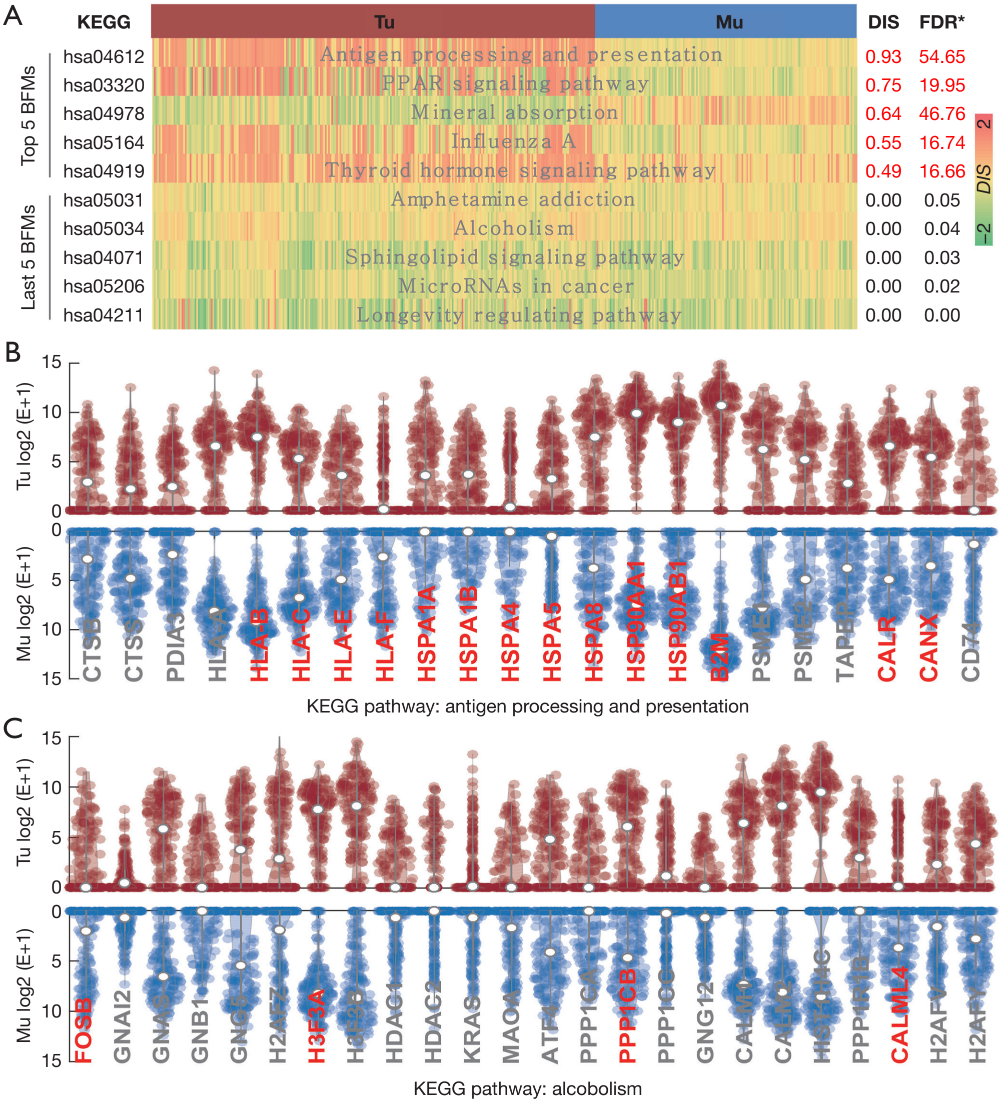

Figure 2 Biological function module analysis of CRC-data. (A) The heatmap shows the IScore of each cell. The top 5 and last 5 BFMs of KEGG pathways were presented to show differences between Tumor (Tu) and normal mucosa (Mu) groups. (B) The violin plot shows gene expression patterns in Tu and Mu groups. Genes displayed in the graph were annotated in the hsa04612 KEGG pathway. (C) The violin plot shows gene expression patterns of the hsa05034 KEGG pathway. Gene names with scarlet font indicate significant differentially expressed genes, and others are represented by the gray font. To better present the expression trend of genes, gene expression values were log-transformed and shown as colored circles for each gene, whereas the empty circles indicate the median of gene expression value across cells. FDR* indicates the log-transformed and converted into the positive value of FDR. KEGG, Kyoto Encyclopedia of Genes and Genomes; BFM, biological function module; FDR, false discovery rate; CRC, colorectal cancer.

population. Despite the existence of a few outliers, the trend remained stable within groups and IScore differences between conditions clearly (*Figure 2A*). Then, results were further analyzed by permutation tests between the 2 conditions and the DISs were quantified by calculating deltas of 2 condition means. A total of 134 KEGG pathways were annotated, wherein, 54% of the pathways were remarkably different between the 2 conditions (P<0.05). However, 8 (5.9%) of the DISs were greater than or equal to the pre-set threshold (0.3), and a majority of DISs (94.1%)

Table 1 The statistical information of the top 5 and last 5 KEGG pathways identified by NonLoss

Accession ID	Function module	Number of genes		Ratio
		Total	Sig	
Top 5 BFMs				
Hsa04612	Antigen processing and presentation	22	14	0.64
Hsa03320	PPAR signaling pathway	13	7	0.54
Hsa04978	Mineral absorption	13	8	0.62
Hsa05164	Influenza A	36	14	0.39
Hsa04919	Thyroid hormone signaling pathway	23	7	0.30
Last 5 BFMs				
Hsa05031	Amphetamine addiction	16	4	0.25
Hsa05034	Alcoholism	24	4	0.17
Hsa04071	Sphingolipid signaling pathway	16	1	0.06
Hsa05206	MicroRNAs in cancer	31	8	0.26
Hsa04211	Longevity regulating pathway	12	2	0.17

Note: 'Total' represents the number of genes annotated to the module and 'Sig' represents the number of significantly differentially expressed genes. The last column is the ratio of 'Sig' and 'Total' column. KEGG, Kyoto Encyclopedia of Genes and Genomes.

showed a difference less than the default. The results accorded with biological principles; sustained homeostasis is the goal of living cell function regulation in the condition of being stimulated by stresses.

For comparison purposes, the top 5 and the last 5 KEGG pathways were used to show differences between each other. The differences of IScore between the 2 conditions were more striking for in the top 5 compared with the last 5 (*Figure 2A*). The differences were closely related with gene expression patterns between 2 conditions. We further found that the ratio of significantly differentially expressed genes in the top 5 KEGG pathways was higher than the last 5 (*Table 1*). These results provided objective evidence for DBMs identified by NonLoss. Notably, since the immune system always fails to destroy tumor cells (25), we found genes of major histocompatibility complex class I emerged with differential expression patterns. Immune-related key genes were significantly downregulated in tumor epithelial cells (B2M, HLA-A, HLA-B, HLA-E, HLA-F; *Figure 2B*). More remarkably, 64% of genes from the hsa04612 pathway were significantly different between the two groups ($P<0.05$). However, only 4 genes showed slight differences between the conditions for hsa05034, which was not DBM (*Figure 2C*). Therefore, the analyses with the KEGG pathway modules showed that NonLoss was able to identify biological significance in particular cancer cell populations.

Differential GO modules identification

A goal of NonLoss was to provide a more robust way to identify significant function modules responding to stimulus. The GO is becoming a more expressive way of describing the function of gene products that allows annotations to be connected together to give a complete function of what each gene does in the context of a larger biological process (26,27).

To test robustness, we further operated on the CRC-data based on GO modules. In total, 386 biological process (BP), 251 molecular function (MF), and 240 cellular component (CC) GO terms were accepted for calculations based on the defaults for each of the parameters. Furthermore, 28/386 (7.2%) BP, 16/251 (6.3%) MF, and 17/240 (7.1%) CC GO terms were identified as DBMs by setting DIS threshold to 0.5 for illustrative purposes (*Figure 3*), which were highly relevant to cancer development, such as "cellular response to hypoxia" in BP, "phospholipid binding" in MF and "specific granule lumen" in CC categories (Figure S1). We found most of the GO terms were not significantly different or had low DIS (*Figure 3*). These results provided further evidence for the biological law which keeps cell state relatively stable, maintaining basic cellular metabolism and

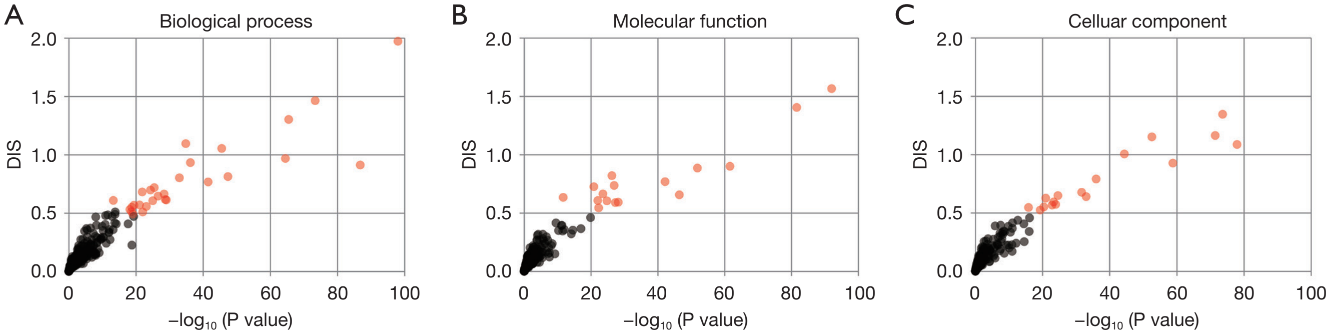

Figure 3 Representative dot plot of DIS (x-axis) versus log-transformed P value (y-axis) analysis of three GO categories. The red spots indicate the GO terms with significant differenced (P<0.05 and DIS ≥0.5) and the black spots indicate GO terms with no major differences (P≥0.05 or DIS <0.5) between conditions. DIS, difference of IScore; GO, Gene Ontology.

promoting survival.

Evaluating robustness of the NonLoss method

To evaluate the robustness of the NonLoss method, random sampling tests were carried out on CRC-data. We first acquired the top 15 GO terms from the BP category, which were the analyses from half the cells of each condition, as the standards for reference. The sampling number (SN) was settled from 50 to 100 and incremented by 10. Next, the number of SN cases were randomly sampled from the rest of the cells from each group 100 times and analyzed by NonLoss with the default values. Therefore, the number of GO term repeats were recorded and the repeat rates were obtained by dividing by 100. By increasing SN from 50 to 100, the repeat rate of each GO term was markedly elevated (*Figure 4A*). For example, when the number of SN increased, the repeat rate of "GO:0071456" went from 0.62 to 0.95. In summary, NonLoss illustrated the variability of scRNA-seq data by taking into account the biological stability or technique. It also displayed the powers of robustness and resistance to the low-quality data.

Furthermore, we introduced the conformance testing for random sample sets. The overlap rates were calculated by taking the intersections of DBMs from random sample sets with a specific SN (*Figure 4B*) from normal mucosa and tumor cell populations, respectively. The average overlap rate significantly increased from 0.27 (SN =50) to 0.68 (SN =100). In fact, this was exactly what our results demonstrated; there may be some small or inconspicuous biological functional differences which were notoriously prone to be enriched or falsely ignored between conditions. This would ultimately lead to poor overlap rates. However, alternatively, the repeat rates of the top GO terms with higher DIS exhibited remarkably high values (for instance, the repeat rate of "GO:0071456" was 0.95, when SN was 100). This was further evidence that it had a high detectability for key biological function sets and a better robust system based on more complete data. This ability of biological feature identification was further demonstrated by a converse-solving strategy. A random and grouped method was used to randomly divide cells from the CRC-data into 2 groups. Then we used NonLoss analysis on these 2 randomly sampled groups. The results were then compared with normal paired groups. We got a significantly (P<0.05) lower overlap rate compared with the normal paired groups from CRC-data (*Figure 4B*). The overlap rates of the random groups for all SN were less than 0.1 and significantly decreased from 0.036 (SN =50) to 0.003 (SN =100). These findings suggested that NonLoss can extract real biological features from different cell populations.

NonLoss highlights blood transcriptional module implicated in MAIT cell activation

Different DE analysis methods can draw different gene lists; however, the overlaps of genes between the methods are usually relatively low (<70%) (28). The dynamic fluctuations in DE gene lists can greatly impact gene enrichment analyses (10,29), leading to enormous deviations in the discovery of biological functions. To compare and verify the ability of function identification reported by NonLoss objectively, we built an entire application on MAIT-data from MAST test data directly and evaluated the outcomes by result comparison (6). A total of 24 blood transcriptional module (30) DBMs were obtained as the default (Table S1).

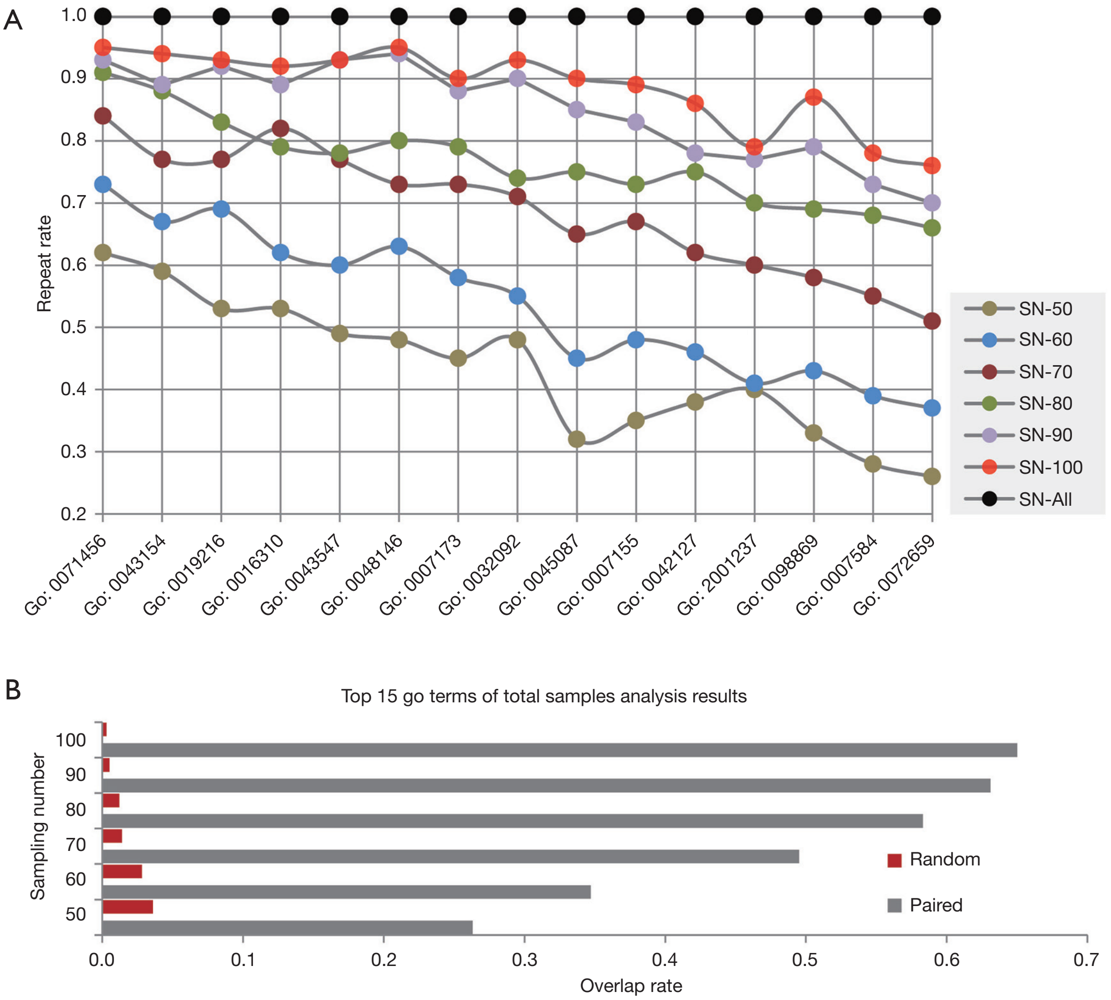

Figure 4 Robustness evaluation of the NonLoss method. (A) Repeat rates of the top 15 GO terms show the frequency of their occurrence for different SN, when randomly sampled 100 times. (B) Overlap rate was calculated by taking the intersection of DBM results of a random sample set with a specific SN. The x-axis shows the average overlap rate of DBMs from random sampling 100 times. The y-axis displays the number of cells randomly sampled from datasets. The gray bars represent 2 groups of the calculation that were randomly sampled from normal mucosa and tumor cell populations, respectively (paired), whereas the crimson bars represent 2 groups of the calculation that were randomly sampled from the CRC-data (random). GO, Gene Ontology; SN, sampling number; DBM, differential biological module.

The intersection of significant function modules of the 2 methods was 11 (45.8%). The major responsive modules were consistent with each other such as "suppression of MAPK signaling", "AP-1 transcription factor network", "spliceosome", "proteasome", and "cell cycle and growth arrest", whereas the NonLoss methods revealed more modules related to T-cell signatures ["signaling in T cells (I)", "enriched in T cells (II)"," T cell differentiation (Th2)" and "mitotic cell cycle in stimulated CD4 T cells"], exclusively. Furthermore, "type I interferon response" (31,32), "phosphatidylinositol signaling system" (33), and "myeloid, dendritic cell activation via NFκB" (34) were consistent with previous findings. Moreover, genes in DBMs exhibited significant differences in expression patterns between the 2 conditions (Figure S2). The results showed that our proposed method was better for presenting the change of biological functions compared with traditional DE methods.

Conjoint analysis of bulk and single-cell RNA-seq data for time-course profiles

As mentioned above, bulk RNA-seq quantified gene expression information on the average level of a cell population showed the general gene expression patterns. It could possess relative stable results from both experimental and biological perspectives. In contrast, scRNA-seq could provide unparalleled resolution to study cellular heterogeneity. However, due to biological variability of individual cells as well as the presence of technical limitations of scRNA-seq, these factors are likely to skew the results. Therefore, we further extended our method to a set of data (GSE52529) with both bulk and single-cell RNA-seq profiles (24).

The data came from a primary human skeletal muscle myoblast (HSMM-data) under high-mitogen conditions (GM) and induced differentiation by switching to low-serum medium (DM) at 4 time points, which has been described in the Methods section. We carried out NonLoss analysis between the cells and samples from different time points (cells collected at 0, 24, 48, and 72 h) for both data sets on the default parameters. Some of the entities we ascertained were worth meditative which may contribute to discovering universal biological laws.

Firstly, there were both intersections and differences of DBMs identified by bulk and scRNA-seq data, respectively. The intersections resulted in 0 *vs.* 24, 0 *vs.* 48, and 0 *vs.* 72 comparisons of 2 technology solutions having 23, 46, and 49 GO BP terms, respectively (Figure S3). The results showed both commonness and specialty on the discovery of biological function changes from bulk RNA-seq and scRNA-seq data. The majority of the terms were associated with cell proliferation and differentiation, such as "negative regulation of G0 to G1 transition", "sister chromatid cohesion", "negative regulation of growth", and so on. Due to bulk RNA-seq data providing more complete information and incorporating more genes for biological modules, bulk RNA-seq data garnered more DBMs under the same conditions and revealed higher stability and repeatability through IScore analyses (*Figure 5A*, available online: https://cdn.amegroups.cn/static/public/atm-21-6401-1.xlsx). Nevertheless, scRNA-seq data was more sensitive to biological features relative to bulk RNA-seq and showed heterogeneity due to unsynchronized development among cell population, or contamination by other cells. For example, the GO terms such as "response to ischemia", "cellular response to glucose starvation" and "response to cadmium ion" reflect the changes of microenvironments from the switch to low-serum medium (DM) that were highly consistent with expectations and were exclusively identified by the scRNA-seq data (available online: https://cdn.amegroups.cn/static/public/atm-21-6401-1.xlsx).

Secondly, by comparing the 0, 24, 48, and 72 h successive time orders using the 2 data types, we found that the DBM overlaps between successive time pairs were considerably smaller (*Figure 5B*). It also revealed that cell physiology states corresponding to stimulations changed at different time points. Finally, with time, physiological cell states gradually became comparatively balanced depending on the condition. This presented as a slowdown in DBM accumulation from both bulk and single-cell data. However, they still exhibited distinct functional variations between cell populations at successive time points (*Figure 5C*).

By comparing DBMs identified from the scRNA-seq and bulk RNA-seq profiles, we could uncover functional changes that were undetectable when averaging over the cell population. In addition, we gained the ability to assess the level of heterogeneity of key differentiation regulators. Further development of this analytical technique may enable us to assess variations in proliferation and differentiation potential across individual cells. Collectively, these findings indicated that NonLoss could identify significant DBMs effectively and could be applied to bulk RNA-seq profiles.

Discussion

Most studied specimens of bulk RNA-seq consist of mixed cell types that display differences in their transcriptomic profiles, as changes in genes which are sensitive or responding to the condition may alter their expression significantly. The situation is even more serious in cancer research which obscures the signatures of tumor characteristics because the intra-tumor heterogeneity of tumor cells is different from normal control. To eliminate such cell type specific effects and secure research authenticity, the scRNA-seq technique can be used to identify rare but significant biological functions between conditions.

Traditional strategies for biological function research have focused on DE gene analysis between conditions of interest and then conducting gene enrichment analysis. To interpret the DE analysis result, GO terms or other biological functional modules have been used to assess the over-representation of a function in the DE genes (gene

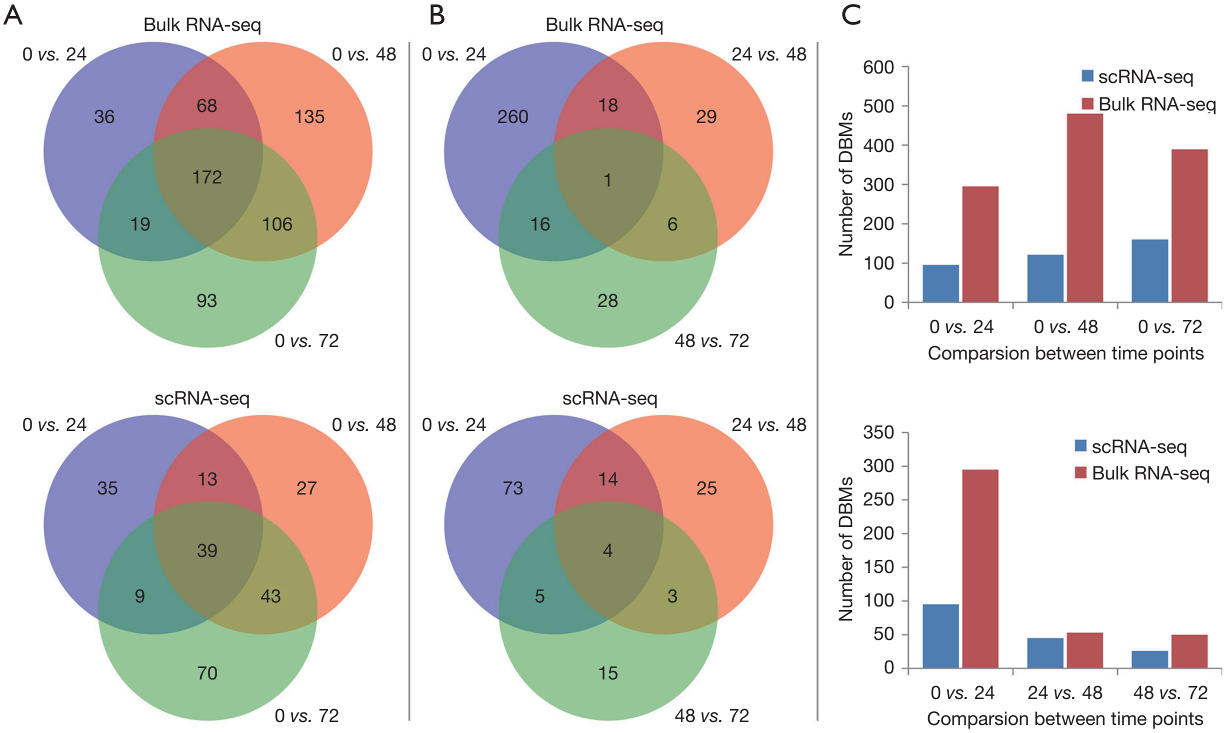

Figure 5 The overlap analysis of DBMs between different time pair comparisons. (A) The Venn diagram shows the number of DBMs overlap between cells of DM and cells collected 24, 48 or 72 h after serum switch, respectively. Upper panel is for bulk RNA-seq data and lower panel is for scRNA-seq data. (B) The Venn diagram shows the number of DBMs overlap between successive time paired comparisons. Upper panel is for bulk RNA-seq data and lower panel is for scRNA-seq data. (C) Number of DBMs histogram statistics. Upper panel is number of DBMs overlap between cells of DM and cells collected 24, 48, or 72 h after serum switch, respectively. The lower panel is number of DBMs overlap between successive time paired comparisons. DBM, differential biological module.

enrichment analysis). Although useful, these methods ignore a lot of information that may provide more robust information about subtle changes at a transcriptional level (FC <2).

It is important to consider biases such as incomplete knowledge of the target genome or functional annotation bases. Our ultimate goal was not to evaluate the results analyzed by DE and gene enrichment methods, but rather to examine whether the most significant features can be more effectively identified through a different analytical approach. Our proposed method, NonLoss, is particularly useful for efficiently translating scRNA-seq transcriptome to biological discoveries from an entire network of genes.

An advantage of the single-cell approach is that we can study the distribution of expression levels across the whole cell population, thereby capturing cell-to-cell variability in gene expression. The expression relationship of genes within the same cell is key to the biological function analysis in our model. Therefore, different calculation methods of expression values do not affect the results. To explore the gain or loss of biological function by different stimuli in a single-cell, we carried out DBM analyses of normalized FPKM or RPKM of scRNA-seq profiles. The higher noise, technical problems, and even stochastic nature of transcription make the interpretation of results and cellular differences difficult to discern. More importantly, DE of a gene may be remarkably different between cells, despite not being functionally critical—and vice versa. Hence, the vast majority of genes with low or no differences should not be omitted from further analyses. Functional analysis was carried out using a small subset of genes called "DE genes". These incomplete data inevitably skewed the results. To account for the confounding factor of expression level, we developed a more robust differential biological functions identification method. We aggregated a functionally related set of genes into one basket and requested the algorithm to re-classify its weight. Therefore, dimensionality reduction from a group of related genes transformed the functional module from a high-dimensional space into a low-dimensional one. This modified module was much easier to visualize and interpret. NonLoss made it possible to acquire faint signals of gene expression and capture miniscule variations between cells or different conditions. When a composite of weak evidence was aggregated, the significant biological functions could be revealed. Furthermore, only the major sample IScores of the function module that were different between conditions were considered significant. In this way, we were able to obtain more reliable function modules with directionality.

Conclusions

NonLoss was able to capture subtle gene expression disturbance in a functional gene-set. This is an optimization that was introduced in scRNA-seq data analysis, which enables biofunctional discovery to optimize DE analysis as a whole rather than as several independent genes. This methodology can more objectively and accurately be used to find important pathways or functions in various diseases and cell conditions. The GOs terms and KEGG pathways as well as self-defined function modules can be used as functional features with NonLoss. To the best of our knowledge, we are the first to propose this type of analytical method for DBM identification, based on whole single RNA sequencing data. Our tool was programmed in Python and is user-friendly, powerful, and easily accessible.

Acknowledgments

We would like to thank all members of Jin Laboratory for discussions, suggestions, and support. We are grateful to Dr. Ting Shao and Dr. Yun Xiao for their comments on an earlier draft of our manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (81871415, to YJ; and 81802428, to WN); the Program for Heilongjiang Postdoctoral Science Foundation (LBH-Z17106; to WN); and the Heilongjiang Touyan Innovation Team Program (to YJ).

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at https://dx.doi.org/10.21037/atm-21-6401

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/atm-21-6401). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Korthauer KD, Chu LF, Newton MA, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol 2016;17:222.
2. Lun ATL, Calero-Nieto FJ, Haim-Vilmovsky L, et al. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. Genome Res 2017;27:1795-806.
3. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015;16:133-45.
4. Peng J, Sun BF, Chen CY, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res 2019;29:725-38.
5. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods 2018;15:255-61.
6. Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 2015;16:278.
7. Jaakkola MK, Seyednasrollah F, Mehmood A, et al. Comparison of methods to detect differentially expressed genes between single-cell populations. Brief Bioinform

2017;18:735-43.
8. Liang S, Ma A, Yang S, et al. A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis. Comput Struct Biotechnol J 2018;16:88-97.
9. Islam S, Zeisel A, Joost S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 2014;11:163-6.
10. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
11. Van Esser JW, Sturm AW. Antimicrobial therapy and temperature. J Antimicrob Chemother 1990;25:716.
12. Delmans M, Hemberg M. Discrete distributional differential expression (D3E)--a tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics 2016;17:110.
13. Wang T, Nabavi S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods 2018;145:25-32.
14. Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Brief Bioinform 2019;20:1583-9.
15. Sun H, Wen X, Li H, et al. Single-cell RNA-seq analysis identifies meniscus progenitors and reveals the progression of meniscus degeneration. Ann Rheum Dis 2020;79:408-17.
16. Gene Ontology C, Blake JA, Dolan M, et al. Gene Ontology annotations and resources. Nucleic Acids Res 2013;41:D530-5.
17. The Gene Ontology C. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47:D330-D8.
18. Ilicic T, Kim JK, Kolodziejczyk AA, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol 2016;17:29.
19. Rasche A, Lienhard M, Yaspo ML, et al. ARH-seq: identification of differential splicing in RNA-seq data. Nucleic Acids Res 2014;42:e110.
20. Li H, Courtois ET, Sengupta D, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet 2017;49:708-18.
21. LESNE A. Shannon entropy: A rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. Mathematical Structures in Computer Science 2014;E240311. .
22. Zhang Y, Liu H, Lv J, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy. Nucleic Acids Res 2011;39:e58.
23. Spearman's rank correlation test. J Clin Nurs 1999;8:763.
24. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381-6.
25. Steinert G, Scholch S, Niemietz T, et al. Immune escape and survival mechanisms in circulating tumor cells of colorectal cancer. Cancer Res 2014;74:1694-704.
26. Gene Ontology C. Gene Ontology Consortium: going forward. Nucleic Acids Res 2015;43:D1049-56.
27. Thomas PD. The Gene Ontology and the Meaning of Biological Function. Methods Mol Biol 2017;1446:15-24.
28. Wang T, Li B, Nelson CE, et al. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics 2019;20:40.
29. Wang X, Cairns MJ. Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. BMC Bioinformatics 2013;14 Suppl 5:S16.
30. Li S, Rouphael N, Duraisingham S, et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat Immunol 2014;15:195-204.
31. Curran E, Chen X, Corrales L, et al. STING Pathway Activation Stimulates Potent Immunity against Acute Myeloid Leukemia. Cell Rep 2016;15:2357-66.
32. Harandi AM, Svennerholm B, Holmgren J, et al. Interleukin-12 (IL-12) and IL-18 are important in innate defense against genital herpes simplex virus type 2 infection in mice but are not required for the development of acquired gamma interferon-mediated protective immunity. J Virol 2001;75:6705-9.
33. Sun Y, Dandekar RD, Mao YS, et al. Phosphatidylinositol (4,5) bisphosphate controls T cell activation by regulating T cell rigidity and organization. PLoS One 2011;6:e27227.
34. van de Laar L, van den Bosch A, van der Kooij SW, et al. A nonredundant role for canonical NF-kappaB in human myeloid dendritic cell development and function. J Immunol 2010;185:7252-61.

Cite this article as: Zhao H, Guo Y, Ma Y, Chen Y, Sun H, Sun D, Wu N, Jin Y. NonLoss: a novel analytical method for differential biological module identification from single-cell transcriptome. Ann Transl Med 2021;9(24):1788. doi: 10.21037/atm-21-6401

Table S1 Blood transcriptional modules (BTM) identified by NonLoss and the overlap comparison to MAST method

BTM Module	DIS	log(FDR)	MAST
type I interferon response (M127)	1.78	29.68	No
suppression of MAPK signaling (M56)	1.57	29.68	Yes
AP-1 transcription factor network (M20)	1.50	38.37	Yes
signaling in T cells (I) (M35.0)	1.26	29.68	No
complement activation (I) (M112.0)	1.07	29.68	No
enriched in B cells (V) (M47.4)	1.07	29.68	No
spliceosome (M250)	0.94	29.68	Yes
phosphatidylinositol signaling system (M101)	0.87	29.68	No
myeloid, dendritic cell activation via NFkB (II) (M43.1)	0.72	29.68	No
enriched for TF motif TNCATNTCCYR (M232)	0.71	29.68	No
enriched in T cells (II) (M223)	0.68	29.68	No
proteasome (M226)	0.63	29.68	Yes
cell cycle and growth arrest (M31)	0.62	29.68	Yes
leukocyte activation and migration (M45)	0.56	29.68	No
transcription elongation, RNA polymerase II (M234)	0.56	29.68	Yes
respiratory electron transport chain (mitochondrion) (M238)	0.54	29.68	Yes
myeloid, dendritic cell activation via NFkB (I) (M43.0)	0.54	29.68	No
translation initiation factor 3 complex (M245)	0.53	29.68	Yes
respiratory electron transport chain (mitochondrion) (M219)	0.52	29.68	Yes
respiratory electron transport chain (mitochondrion) (M216)	0.49	29.68	Yes
cell cycle, ATP binding (M144)	0.47	29.68	Yes
leukocyte differentiation (M160)	0.44	29.68	No
T cell differentiation (Th2) (M19)	0.43	38.37	No
mitotic cell cycle in stimulated CD4 T cells (M4.9)	0.42	38.37	No

BTM, Blood transcriptional modules; DIS, difference of IScore; FDR, false discovery rate.

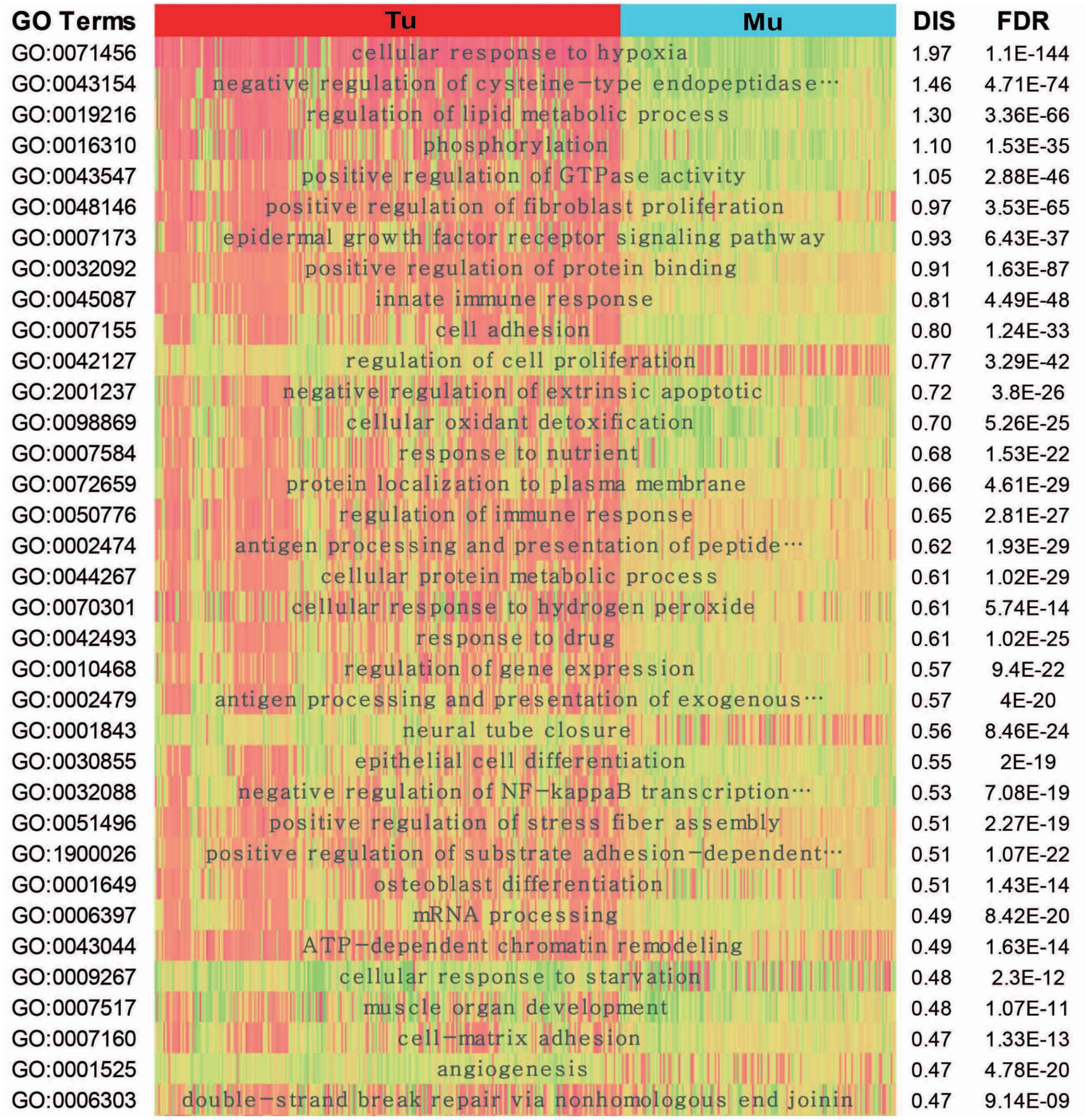

Figure S1 Heatmap shows IScore difference of GO BP category between Tumor and Normal mucosa. Furthermore, 28/386 (7.2%) BP GO terms were identified as DBMs by setting DIS threshold to 0.5. IScore, integrated scores; DBMs, differential biological modules; GO, Gene Ontology; BP, biological process; DIS, difference of IScore.

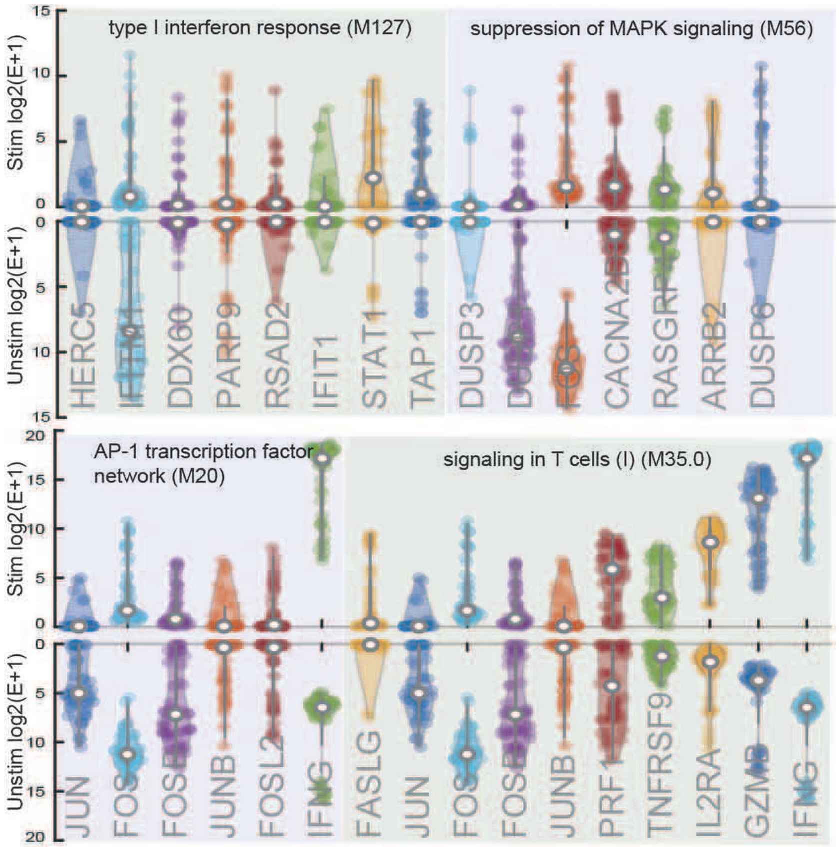

Figure S2 The violin plot shows gene expression patterns in non-stimulated (Unstim) and cytokine-stimulated (Stim) MAST cell. The genes displayed in the graph are annotated in "type I interferon response (M127)", "suppression of MAPK signaling (M56)", "AP-1 transcription factor network (M20)" and "signaling in T cells (I) (M35.0)". The letter (E) indicates FPKM value.

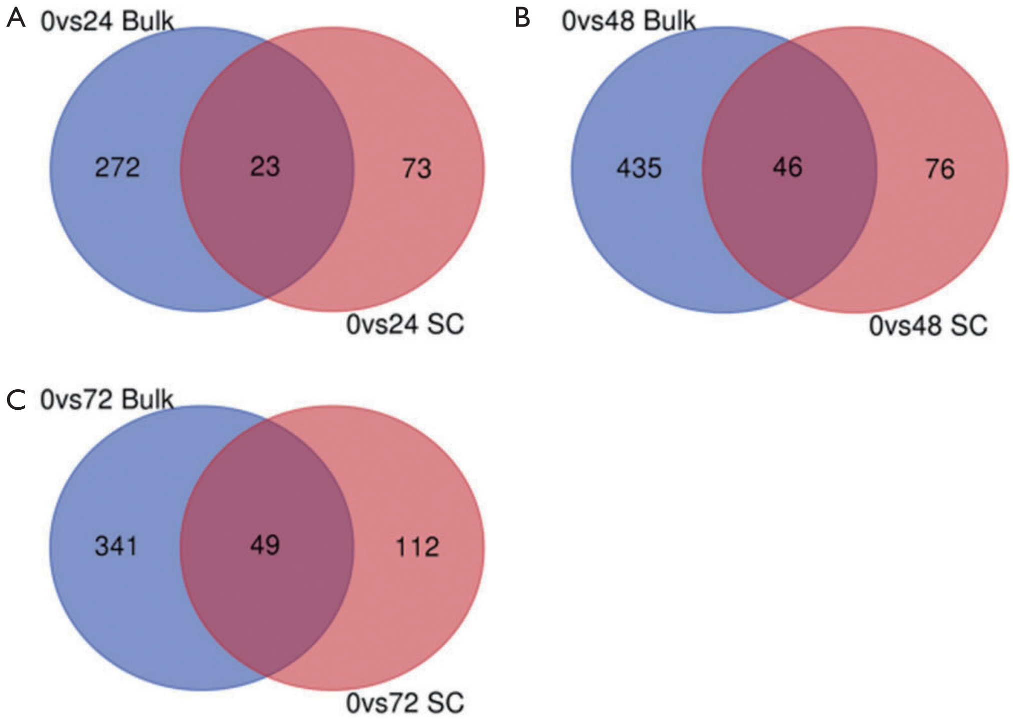

Figure S3 The venn diagram shows the number of DBMs overlap between scRNA-seq and Bulk RNA-seq data which were collected at different time points. DBMs identified by NonLoss between any pair of time points using bulk RNA-seq and single cell RNA-seq data respectively. (A) Calculated by cells of DM and cells collected 24h after serum switch. (B) Calculated by cells of DM and cells collected 48h after serum switch. (C) Calculated by cells of DM and cells collected 72h after serum switch.