Computers in Biology and Medicine 157 (2023) 106733

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Gene function and cell surface protein association analysis based on single-cell multiomics data

Huan Hu [a,b,c,1], Zhen Feng [d,1], Hai Lin [c], Jinyan Cheng [c], Jie Lyu [c], Yaru Zhang [e], Junjie Zhao [c,f], Fei Xu [a,c], Tao Lin [g], Qi Zhao [h,*], Jianwei Shuai [a,b,c,g,**]

[a] Department of Physics, Fujian Provincial Key Laboratory for Soft Functional Materials Research, Xiamen University, Xiamen, 361005, China
[b] National Institute for Data Science in Health and Medicine, State Key Laboratory of Cellular Stress Biology, Innovation Center for Cell Signaling Network, Xiamen University, Xiamen, 361005, China
[c] Wenzhou Institute and Wenzhou Key Laboratory of Biophysics, University of Chinese Academy of Sciences, Wenzhou, 325001, China
[d] First Affiliated Hospital of Wenzhou Medical University, Wenzhou Medical University, Wenzhou, 325000, China
[e] Institute of Biomedical Big Data, School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou, 325027, China
[f] Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, 510000, China
[g] Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), Wenzhou, 325001, China
[h] School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China

ARTICLE INFO

Keywords:
Single-cell
Multiomics
Cell surface protein
Association analysis
Computing framework

ABSTRACT

Single-cell transcriptomics provides researchers with a powerful tool to resolve the transcriptome heterogeneity of individual cells. However, this method falls short in revealing cellular heterogeneity at the protein level. Previous single-cell multiomics studies have focused on data integration rather than exploiting the full potential of multiomics data. Here we introduce a new analysis framework, gene function and protein association (GFPA), that mines reliable associations between gene function and cell surface protein from single-cell multimodal data. Applying GFPA to human peripheral blood mononuclear cells (PBMCs), we observe an association of epithelial mesenchymal transition (EMT) with the CD99 protein in CD4 T cells, which is consistent with previous findings. Our results show that GFPA is reliable across multiple cell subtypes and PBMC samples. The GFPA python packages and detailed tutorials are freely available at https://github.com/studentiz/GFPA.

1. Introduction

Single-cell transcriptomics has revolutionized our understanding of complex biological systems. It enables the creation of a comprehensive cellular map of an organism through routine measurements of gene expression in thousands of individual cells [1–3]. The transcriptome is just one aspect that determines cell type, state, and function, among many other regulatory controls. To further reveal the heterogeneity of cells, single-cell multiomics techniques were developed [4], which refer to the simultaneous measurement of multiple omics information such as transcriptome [2], genome [5], epigenome [6], proteome [7] or spatial location [8] at single-cell resolution.

However, most multiomics analysis frameworks were primarily focused on integrating data from different modalities to create low-dimensional representations. For example, SeuratV4 used a weighted combination of two modalities to define cell states [9]. CITEMO was an efficient method for analyzing single-cell multiomics data by combining multiple modal principal components to quickly estimate multimodal cell representations. This approach provided a new way to uncover the complex relationships between different modalities in single-cell multiomics data [10]. TotalVI trained a neural network to infer the distribution of low-dimensional embeddings of multimodal representations to obtain a noise-free representation of single cell states [11]. The DPI method was based on the fusion of parameters from three separate neural networks, with the goal of inferring the multimodal distribution of data. This approach enabled the integration of information from

* Corresponding author.
** Corresponding author. Department of Physics, and Fujian Provincial Key Laboratory for Soft Functional Materials Research, Xiamen University, Xiamen, 361005, China.
E-mail addresses: zhaoqi@lnu.edu.cn (Q. Zhao), jianweishuai@xmu.edu.cn (J. Shuai).
[1] These authors contributed equally to the paper as first authors.

https://doi.org/10.1016/j.compbiomed.2023.106733
Received 19 January 2023; Received in revised form 8 February 2023; Accepted 28 February 2023
Available online 1 March 2023

multiple data modalities to produce a more comprehensive understanding of the cell system [12]. GLUE used variational graph autocoders to learn the weights of omics features and thus to correlate the low-dimensional embeddings of different omics to represent cellular heterogeneity [13]. Although the aforementioned methods have shed light on cellular heterogeneity to a certain degree, they have yet to fully tap into the potential of multiomics data. The central dogma of genetics suggests that various omics data are interconnected. For instance, combining the analysis of chromatin accessibility (single-cell DNA data) with gene expression profiles (single-cell RNA data) presents a chance to uncover enhancer-driven gene regulatory networks [14,15]. In this study, we concentrated on investigating the correlation between single-cell RNA data and protein data.

Recently, various sequencing technologies have emerged that enable the simultaneous measurement of both gene expression and cell surface protein levels in single cells, including CITE-seq [16], REAP-seq [17], ECCITE-Seq [18], ASPA-Seq [19], DOGMA-seq [19], Total-Seq [16,20], PHAGE-ATAC [21], among others. These technologies offer exciting new opportunities for single-cell multiomics studies. Cell surface proteins play a crucial role in regulating cell-to-cell communication and interaction with the extracellular environment. These proteins are located on the surface of the cell membrane, either spanning it or anchored within it, and serve as key mediators in transmitting signals both into and out of the cell [22–27]. Cell surface proteins perform specific functions at the cell membrane, such as nutrient and ion transport, intercellular interactions, receptor-mediated signal transduction, enzymatic responses, and immune recognition [22,28–30]. In fact, more than 60% of cell surface proteins have been the targets of the approved drugs for human diseases [22,28]. Cell surface proteins have been confirmed the importance in multiple aspects, but they are used as cellular markers in most studies and their potential functions have not been investigated.

In this study, we introduced a novel analysis framework, gene function and protein association (GFPA), for exploring the relationship between cell surface proteins and gene function. Rather than relying on the strength of the association between genes and their corresponding proteins, GFPA infers protein functions from those of associated genes and incorporates P-values into its scoring system to better represent the strength of the associations. Applying GFPA to human peripheral blood mononuclear cells (PBMC), we observed an association between EMT and the CD99 protein in CD4 T cells, which aligns with previous findings. Our results demonstrated the reliability of GFPA across multiple cell subtypes and PBMC samples.

2. Materials and methods

2.1. Single-cell multiomics data

All data in this paper come from a COVID-19 disease progression study conducted by Muzlifah Haniffa et al. [31]. They performed single-cell transcriptome, surface proteome, T and B lymphocyte antigen receptor profiling on more than 780,000 PBMCs from a cross-sectional cohort of 130 COVID-19 patients of varying severity. We extracted a total of 24 healthy PBMC samples. The cell annotations were all from the ref. [31].

Usually, we analyzed a certain cell subtype in the sample, instead of the whole sample. We defined the dataset of cell subtypes as C containing I cells (Eq. (1)), where I depended on the number of cells collected in the experiment. The transcriptome of the dataset, defined as X, was a count matrix containing J genes from I cells (Eq. (2)), where J depended on the sequencing depth of the kit. Similarly, the cell surface proteins of the dataset, defined as P, was a count matrix containing I cells and K types of proteins (Eq. (3)), where K was determined by the pre-designed antibody type. For each cell, transcriptome and cell surface protein data were measured simultaneously (Eq. (4)).

$$C_I = (C_i) \tag{1}$$

$$X_{I*J} = (x_{ij}) \tag{2}$$

$$P_{I*K} = (p_{ik}) \tag{3}$$

$$C_i = (x_i, P_i) \tag{4}$$

2.2. Gene function database

Gene function data were extracted from MSigDB [32,33], a database containing many gene sets from various perspectives such as location, function, metabolic pathway, and target binding. We curated gene function sets for two species, human and mouse, from MSigDB (Supplementary Table 1). The data we download from MSigDB were saved in GMT format. The GMT format is a tab-separated list of gene sets, where each row is a separate gene set. The first column must contain the name of the gene set and the second column is used for a short description. In this study, the large-scale gene function database, defined as S, contained M gene function sets (Eq. (5)) where M was the database selected by the user (Supplementary Table 1). Any element S_m in S was expressed as a set of l types of genes (Eq. (6)).

$$S_M = \{S_1, S_2, \ldots, S_m, \ldots\} \tag{5}$$

$$S_m = \{g_1, g_2, g_3, \ldots, g_l\} \tag{6}$$

2.3. Data preprocessing

2.3.1. Gene expression data preprocessing

It is well known that transcriptome and cell surface proteins have different biological properties and therefore they should be preprocessed by different methods.

Considering that over normalized data may change the data properties, we performed only the simplest logarithmic transformation of X [34]. X was added with 1 to eliminate the negative infinity introduced by 0 during the logarithmic transformation. The normalized single-cell transcriptome matrix $X_{normalization}$ was denoted as follows:

$$X_{normalization} = log(X + 1) \tag{7}$$

2.3.2. Cell surface protein data preprocessing

Previous studies have shown that Tag, which counts protein abundance, may bind non-specifically to cell surface proteins [9,10,16,18,19]. This suggests that cell surface protein data may contain noise. Previous studies have shown that centered logarithmic ratio (CLR) can eliminate noise to some extent [10,16,18]. The normalized cell surface protezain matrix $P_{normalization}$ was shown as follows:

$$P_{normalization} = CLR(P) = \left[ln\left(\frac{p_1}{g(p)}\right), ln\left(\frac{p_2}{g(p)}\right), \ldots, \left(\frac{p_k}{g(p)}\right)\right] \tag{8}$$

2.4. Gene function center

To quantify the utility of gene function, we proposed gene function center. We believed that the key was the concentration of genes, so the expression of genes, also known as transcriptome, needed to be introduced when the gene function center was calculated.

Deriving the gene function center for S_m included three steps. First, taking the genes from X that were consistent with s_m and building the matrix X_{s_m} (Eq. (9)). Next, the sum of X_{s_m} in the gene dimension was calculated (Eq. (10)). Finally, the Euclidean center of X_{s_m} was used as the gene function center (Eq. (11)).

$$X_{s_m} = \left(X_{g_1}, X_{g_4}, \ldots, X_{g_l}\right) \tag{9}$$

$$X_{sum_m} = \sum X_{s_m} \tag{10}$$

$$X_{center_m} = \frac{X_{sum_m}}{l} \tag{11}$$

$X_{centerm}$ can be interpreted as the average of the concentrations of genes contained in s_m. $X_{centerm}$, determined jointly by s_m and its gene expression, quantifies the utility of the gene set.

2.5. Correlation analysis algorithm

We analyzed the association of gene function and cell surface proteins using the correlation algorithm. Here, gene function was quantified using Eq. (11). Before performing the correlation analysis, we used Z-Score scaling (Eq. (12)) to map gene function centers and cell surface proteins to the standard normal space (Eqs. (13) and (14)), respectively.

$$ZScore = \frac{v - \mu}{\sigma} \tag{12}$$

$$X_{z_m} = ZScore(X_{center_m}) \tag{13}$$

$$P_{z_k} = ZScore(P_{normalization_k}) \tag{14}$$

In Eq. (12), v represented the vector, its mean and standard deviation was denoted by μ and σ, respectively. Z-Score scaling eliminated the difference in magnitude between the data to a comparable scale. In Eq. (13), m was the index of gene function set. In Eq. (14), k was the index of cell surface protein types.

In GFPA framework, Pearson, Spearman, and Kendall correlation algorithms were implemented based on the python package "scipy" (1.9.3). By default, we applied the Spearman correlation algorithm to calculate the correlation. It was worth noting that the p-values need to be corrected after multiple tests. Here, we introduced the BH algorithm for p-value correction [35]. We applied the python package "statsmodels" (0.13.5) to implement the BH algorithm. The adjusted p value less than 0.01 (default) was considered reliable.

2.6. GFPA score

Correlation analysis involved a trade-off between correlation values and p-values. On the one hand, a large correlation value means a strong association. On the other hand, only a very small p-value was statistically significant. Usually, researchers can only choose one of the statistics as a ranking criterion. The proposed GFPA score has the ability to consider both correlation and p-values (Eq. (15)).

$$GFPA_{score} = \left(1+\beta^2\right)\frac{|correlation| * (1-p)}{\left(\beta^2 * |correlation|\right) + (1-p)} \tag{15}$$

$$GFPA_{threshold} = \left(1+\beta^2\right)\frac{|correlation_{threshold}| * (1-p_{threshold})}{\left(\beta^2 * |correlation_{threshold}|\right) + \left(1-p_{threshold}\right)} \tag{16}$$

$$reliablity = \begin{cases} 1, GFPA_{score} > GFPA_{threshold} \\ 0, GFPA_{score} < GFPA_{threshold} \end{cases} \tag{17}$$

$$adj\_GFPA_{score} = \left(1+\beta^2\right)\frac{|correlation| * (1-p)}{\left(\beta^2 * |correlation|\right) + (1-p)} * reliablity \tag{18}$$

Considering that the range of correlation was from 1 to −1 and the range of p-value is from 0 to 1, we took the absolute value of correlation to make their data ranges consistent. Notably, the adjusted p-values were derived from multiple tests and the GFPA calculation was about individuals, leading to the inapplicability in GFPA calculation. β defaults to 1, which controlled for *correlation* and $1-p$ preferences. Furthermore, we proposed two criteria to strictly check the reliability of GFPA scores. First, the adjusted p-value must be lower than $p_{threshold}$ (default $p_{threshold} = 0.01$). Second, the GFPA score must be greater than $GFPA_{threshold}$ (Eq. (16)) (default $correlation_{threshold} = 0.5$ and $GFPA_{threshold} = 0.664430$). Only the GFPA scores for gene function and cell surface protein pairs matching these two criteria are reliable (Eq. (17)). adj_GFPA_{score} can integrate *reliablity* into $GFPA_{score}$ (Eq. (18)).

2.7. Gene weight model

We applied the random forest to infer the weight of each gene in the gene set. The input of the random forest model is X_{zm} (Eq. (13)) and the output is P_{zk} (Eq. (14)). We took the importance score of random forest for each input feature as the weight of the gene. A larger importance score represented a greater effect of the gene on a specific cell surface protein. The random forest model was derived from the implementation of the python package "scikit-learn" (1.1.3).

3. Results

3.1. The workflow of GFPA

In this study, we introduced a novel framework for association analysis, referred to as Gene Function and Protein Association (GFPA). GFPA was designed to automatically explore the relationships between gene function and cell surface proteins. Careful consideration of biological factors was incorporated into each step of GFPA, leading to a robust and reliable method. We recommend starting the analysis by first subdividing the cells into subtypes (Fig. 1A). Single-cell sequencing technology has enabled us to overcome the limitations of bulk sequencing by allowing us to measure the status of each cell subtype. The cell subtype data we extract includes both single-cell gene expression and cell surface protein components (Fig. 1B). We summarized a number of gene function descriptions to established a link between gene expression and function descriptions, and called them gene function sets (Fig. 1C). GFPA can transform gene profiles into gene function centers to quantify a specific gene function (Fig. 1D). The gene function center was jointly determined by the gene expression data and the user-selected gene function database. It is worth noting that the user must specify the gene function database according to his/her study. GFPA supported multiple gene function databases in humans and mice (Supplementary Table 1). The choice of gene function database directly determined the type of gene function center. We used GFPA score to measure the association between gene function centers and cell surface proteins (Fig. 1E). The GFPA score combined the utility of correlation and p-value, which can be considered as an indicator for the association between gene function and protein. The higher the GFPA score, the stronger the association is. Specifically, the GFPA score can be visualized as a scatter plot (Fig. 1E). Further, we introduced random forest to analyze the effect of the expression level of each gene on cell surface proteins (Fig. 1E). The influence of each gene in the gene function set was visualized as gene weights (Fig. 1E). In conclusion, we represented a convenient framework for investigating the relationship between gene function and cell surface proteins. GFPA took biological considerations into account in each step, offering insights into gene and protein functions and potentially guiding future biomedical experiments.

3.2. GFPA used gene set as a proxy for gene function

It is necessary to use gene sets rather than individual types of genes as a basis for gene function. In many cases, proteins and their corresponding genes were translated into each other to perform downstream analysis. This process required the assumption of a strong positive correlation between genes and the corresponding proteins. However, we found the assumption was not always satisfied in single-cell multiomics data. This implied that the association of individual types of genes and proteins is not significant and GFPA addressed this situation by using gene sets as a proxy for gene function.

We analyzed a PBMC sample from a healthy contributor (Fig. 2A).

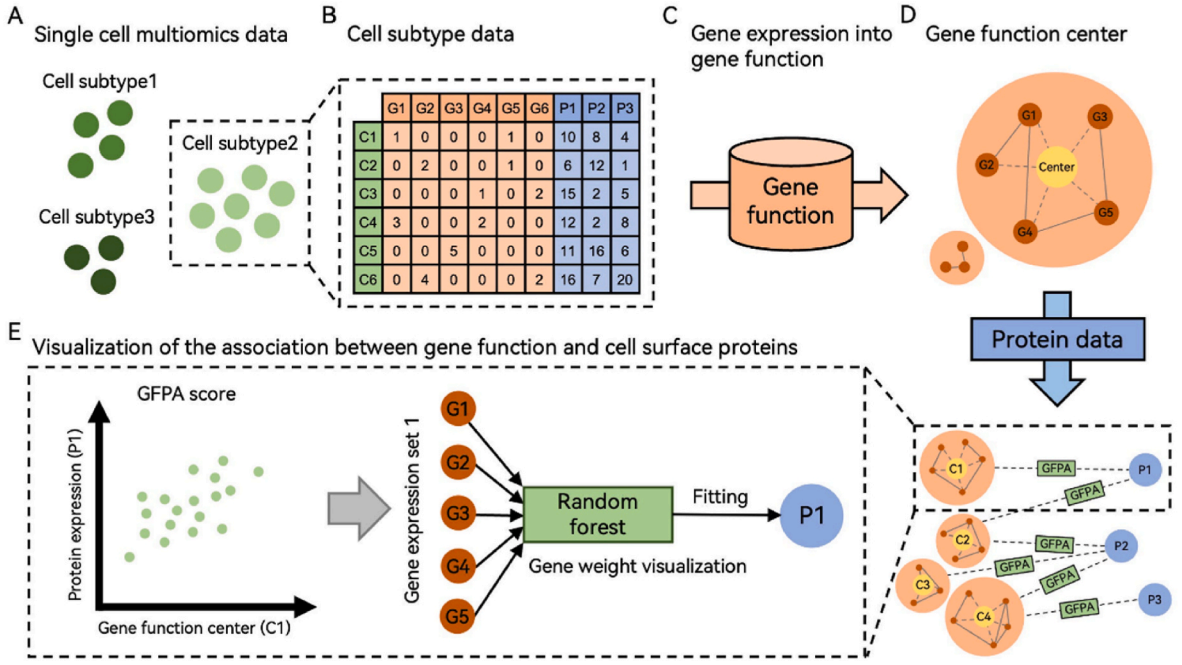


Fig. 1. The workflow of GFPA. First, cell subtypes were extracted from single-cell multiomics data (A and B). Gene expression data in cell subtypes were transformed into gene functions by GFPA (C). We quantified gene function in terms of gene function centers (D). Finally, gene function centers were paired with protein data and their associations were calculated using GFPA (E). Specifically, the association of cell function and cell surface proteins can be visualized with GFPA scores. Further, the GFPA can mine the importance of each gene in gene function for cell surface proteins (E).

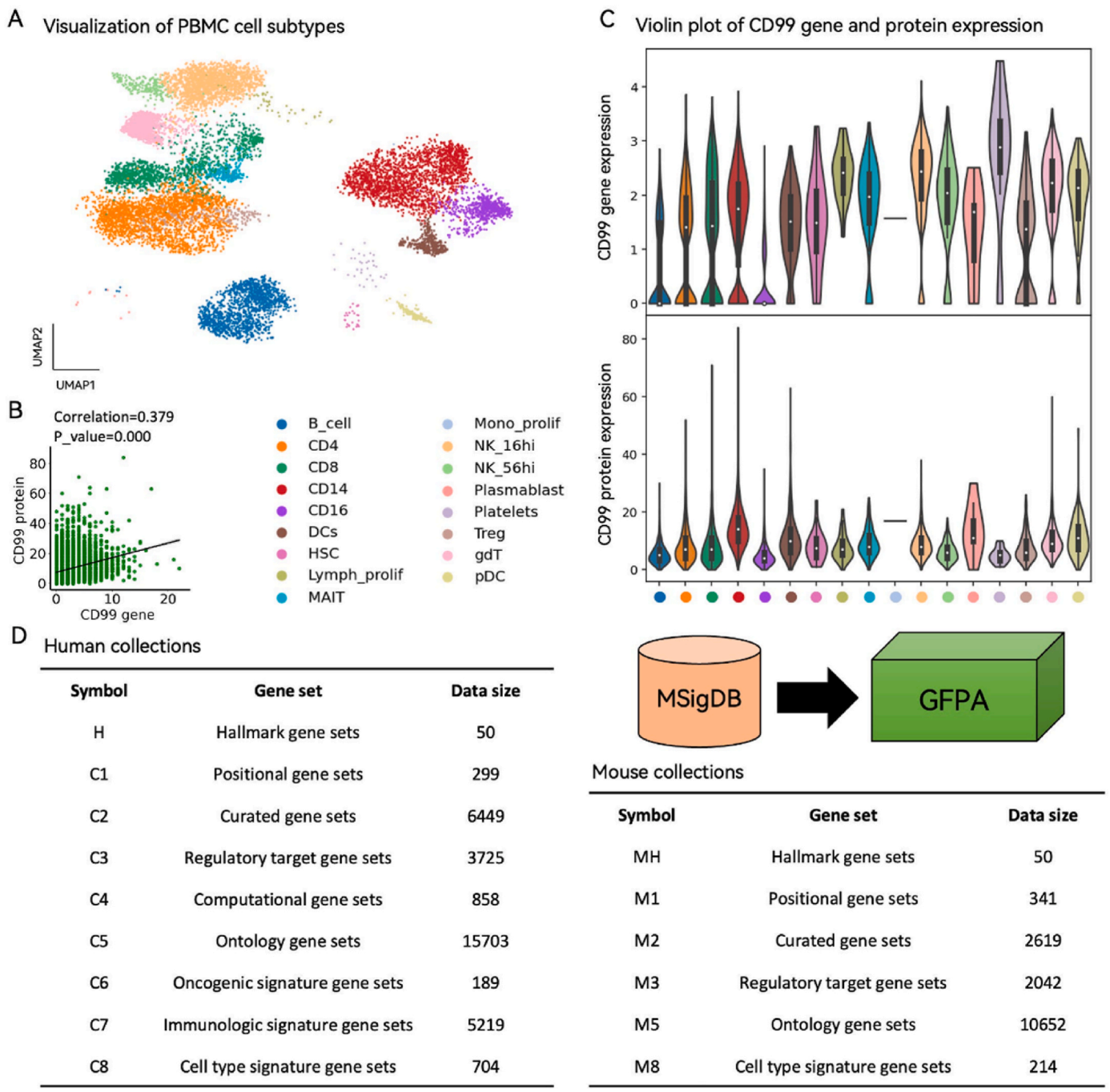
The sample (id: MH8919333) was from a dataset containing 24 PBMC samples and it contains 12,081 cells with 24,737 genes and 192 cell surface protein data. We took CD99 protein as an example to study the association between genes and proteins (Fig. 2B). We can find that they were not significantly correlated. We further explored the expression patterns of CD99 genes and proteins in each cell subtype and found the difference, as shown in Fig. 2C. Concretely, CD99 gene was highly expressed in Plasmablast, but the corresponding protein was not high, which demonstrated that gene data and protein data were not completely equal. In addition, single cell gene expression profiles may suffer from Dropout i.e. low levels of gene deletion [36]. This made individual types of genes (e.g., gene markers [37]) less reliable than those of proteins. Besides, collections of multiple genes with similar functions were more reliable (e.g., enrichment analysis) [38–40]. At the data quality level, gene sets can mitigate the effects of Dropout by using the substitution of other genes in the set for missing values for a single gene. At the gene function level, all genes have multiple functions and genes located in different contexts exhibit different functions. The strategy of conducting single-cell analysis with a gene function perspective was widely employed in many studies [38–42]. Based on the above considerations, GFPA was modeled by a collection of genes rather than individual types of genes. We configured nine human gene collections and six mouse gene collections for GFPA from MSigDB (Fig. 2D and Supplementary Table 1) [32,33]. Users were only required to specify the database according to Symbol, and GFPA will automatically convert the gene expression data into gene set data.

3.3. GFPA can explore the associations between gene function and protein

We took the CD4 T cells in the MH8919333 sample as an example to illustrate the function of GFPA (Fig. 3A). It was required to specify the gene function database before performing GFPA analysis. We selected the human hallmark gene set (Symbol is "H") as the gene function set (Fig. 2D and Supplementary Table 1). GFPA calculated the correlation between each cell surface protein and gene function in CD4 T cells separately. GFPA provided three correlation algorithms: Pearson [43], Spearman [44], and Kendall [45]. In this study, we chose the Spearman correlation coefficient as the metric for all gene collections. GFPA calculated the GFPA score by correlation and p-value and the results were sorted by adjusted GFPA score in descending order (Fig. 3B). To avoid Type I errors [46], we used the adjusted p-values instead (Fig. 3B). The reliability of the results was assessed according to the correlation and the adjusted p-values (Fig. 3B). Results with a GFPA score greater than 0.664430 and an adjusted p-value less than 0.01 were considered reliable. The adjusted GFPA score can further integrate Reliability into GFPA score and the score greater than zero is regarded as reliable. We found that only two gene function and cell surface protein pairs were reliable in CD4 T cells (Fig. 3B). The item with the highest ranked GFPA score showed that CD99 protein was related to the EMT process (Fig. 3C). Several studies have suggested that CD99 protein can serve as a phenotype for EMT [47–49]. Our results were consistent with previous studies [47–50]. Furthermore, we evaluated the association of genes in EMT with CD99 proteins and we found a CTHRC1 gene with a significant association (Fig. 3D). The protein corresponding to the CTHRC1 gene has been reported to be a cancer-associated protein related to multiple signaling and tumor metastasis [51]. Previous studies showed that CTHRC1 can upregulate the expression of EMT-related markers while CD99 has been shown to act as a marker for EMT [51–53]. This evidence implied a possible association between CTHRC1 and CD99. In conclusion, the above results demonstrated that GFPA has the ability to reveal the association between gene function and cell surface proteins, and it helped researchers to unravel the mechanisms of disease development.

3.4. GFPA was a robust analysis framework

To test the robustness of GFPA, we analyzed the associations between gene function and cell surface protein in CD8 T cells. CD8 T cells and CD4 T cells should have similar GFPA analysis results since they both belong to T cells. We found a reliable association between EMT and CD99 proteins in the results of GFPA analysis from CD8 T cells (Supplementary Figs. 1A and 1B). Furthermore, we also identified an association of the CTHRC1 gene in EMT with CD99 in CD8 cells



D Human collections

Symbol	Gene set	Data size
H	Hallmark gene sets	50
C1	Positional gene sets	299
C2	Curated gene sets	6449
C3	Regulatory target gene sets	3725
C4	Computational gene sets	858
C5	Ontology gene sets	15703
C6	Oncogenic signature gene sets	189
C7	Immunologic signature gene sets	5219
C8	Cell type signature gene sets	704

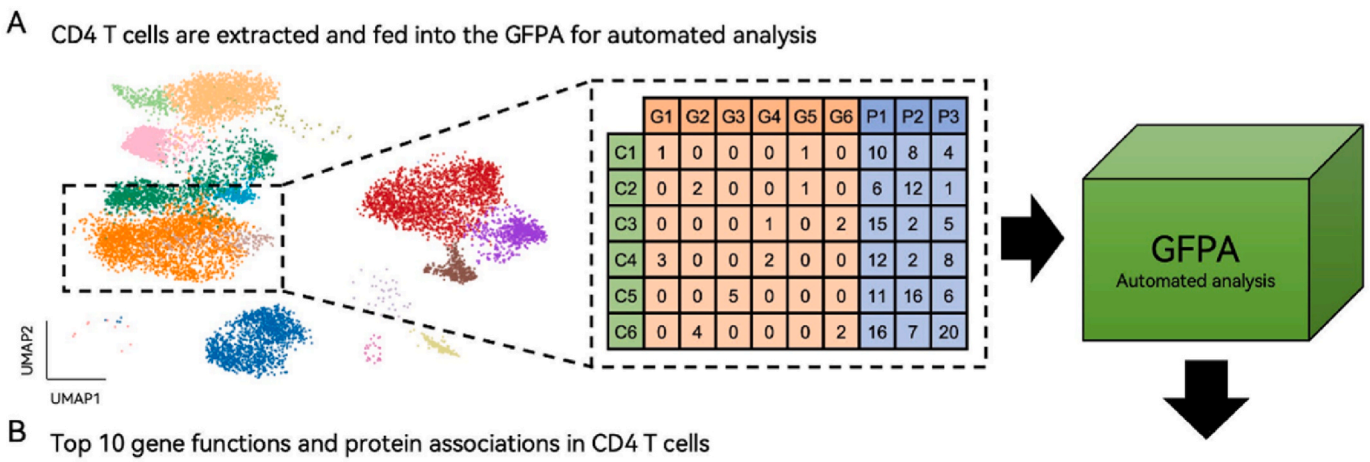
Mouse collections

Symbol	Gene set	Data size
MH	Hallmark gene sets	50
M1	Positional gene sets	341
M2	Curated gene sets	2619
M3	Regulatory target gene sets	2042
M5	Ontology gene sets	10652
M8	Cell type signature gene sets	214

Fig. 2. We analyzed all cells in an annotated PBMC sample (A) for correlation of CD99 gene and protein (B). Each cell subtype's CD99 genes and proteins were visualized by violin plots (C). The GFPA was configured with the human and mouse gene sets collected from MSigDB (D).

(Supplementary Fig. 1C) [54]. These results were consistent with those of CD4 T cells. Further, we performed GFPA analysis on all cell subtypes in MH8919333. Considering the low reliability of statistics with data number less than 30, only cell subtypes with cell number more than 30 were analyzed. We found the reliable association of EMT and CD99 only in T cells (Table 1). Previous studies have shown that CD99 was closely associated with immunotherapeutic T cells using CAR T cell therapy [49, 50]. This also suggested that the association between the EMT and CD99 identified by GFPA was reliable.

We further checked the general applicability by performing GFPA analysis within 24 PBMC samples. CD4 T cells from 24 PBMC samples were extracted and validated for EMT and CD99 associations. We found reliable EMT and CD99 associations for CD4 T cells in most of the samples (Table 2). Among them, MH8919227, MH8919226 and BGCV01_CV0902 have less than 30 CD4 T cells and they were not involved in GFPA analysis (Table 2). In total, only 5 results out of 21 analyses were considered unreliable. Even the results were considered unreliable, four out of them (BGCV15_CV0944, BGCV13_CV0934, BGCV09_CV0917 and BGCV02_CV0902) revealed a weak positive association between EMT and CD99. These results imply that the association between EMT and CD99 proteins is universal. In addition, it also indicates that the GFPA analysis is not affected by sample batches and it is reliable.



B Top 10 gene functions and protein associations in CD4 T cells

Gene function	Protein	Correlation	P_value	Adj_p_value	Reliablity	GFPA_score	Adj_GFPA_score
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	**CD99**	**0.569**	**$1.76x10^{-198}$**	**$1.69x10^{-194}$**	**1**	**0.725**	**0.725**
HALLMARK_HYPOXIA	CD99	0.519	$1.15x10^{-159}$	$5.53x10^{-156}$	1	0.683	0.683
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	CD45RA	-0.492	$1.99x10^{-141}$	$6.37x10^{-138}$	0	0.660	0
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	ITGA6	0.492	$3.20x10^{-141}$	$7.68x10^{-138}$	0	0.660	0
HALLMARK_APOPTOSIS	CD99	0.482	$7.81x10^{-135}$	$1.50x10^{-131}$	0	0.651	0
HALLMARK_TNFA_SIGNALING_VIA_NFKB	CD99	0.479	$1.23x10^{-132}$	$1.97x10^{-129}$	0	0.647	0

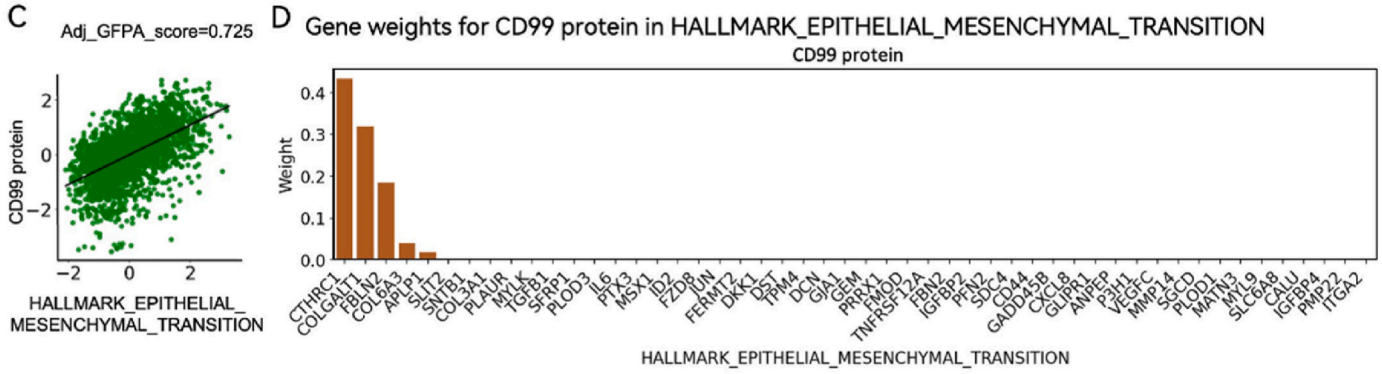


Fig. 3. GFPA analysis of gene function and cell surface protein associations in specific cell subtypes. CD4 T cells were extracted from PBMC samples to perform GFPA analysis (A). The results of GFPA analysis (B) and specific associations (C) were visualized. GFPA can infer the importance of each gene in a gene function collection for a specific protein (D).

4. Discussion and conclusion

In this study, we developed a framework named GFPA to explore gene function and protein associations in single-cell multiomics data. We applied GFPA to successfully identify an association between EMT gene function and CD99 protein. We found this association in both CD4 T and CD8 T cells from the same PBMC sample. In addition, we obtained this result in other PBMC samples. Many previous works have shown the association between EMT and CD99. These results not only demonstrated the ability of GFPA to explore gene function and protein associations, but also indicated its robustness.

We believed that the success of GFPA can be attributed to several factors. First, the gene sets and cell surface proteins analyzed by GFPA were derived from the same cell subtype. Single-cell multiomics experiments allow for a combined assessment of gene expression and cell surface proteins. In one cell subtype, there were no non-biological differences such as batch effects between transcriptome and cell surface protein data, which was inevitable in Bulk-seq. Single-cell multiomics sequencing data was more reliable than Bulk-seq data which guaranteed the reliable GFPA analysis results. Second, GFPA used gene sets rather than individual genes and proteins for association analysis. Dropout events result in missing data for some single-cell gene profiles. Gene sets can properly mitigate the effects of Dropout for individual types of genes. In addition, unlike proteins, the individual types of genes were difficult to have a direct access to complex biological functions. Gene sets were more suitable for describing the function of cells than individual types of genes. Finally, the framework of GFPA, where the score unified correlation and p-value, sets strict evaluation criteria. Since a very small p-value caused a large GFPA score, we set the $GFPA_{threshold}$ as 0.664430 rather than 0.5, and specify that GFPA scores greater than $GFPA_{threshold}$ were likely to be reliable. In addition, we noted that multiple calculations of correlations may lead to Type I errors. We introduced the Benjamini & Hochberg (BH) algorithm to adjust the p-value and only results with GFPA scores greater than $GFPA_{threshold}$ and adjusted p-values less than $p_{threshold}$ (default $p_{threshold} = 0.01$) were considered to be reliable. These strict strategies further guarantee the reliability of

Table 1
GFPA analysis of EMT and CD99 proteins in each cell subtype from MH8919333 samples. GFPA score greater than 0.664430 and an adjusted p-value less than 0.01 were considered reliable.

Celltype	Cell count	Correlation	P_value	Adj_p_value	Reliablity	GFPA_score	Adj_GFPA_score
CD4	2305	0.569	0.000	0.000	1	0.725	0.725
CD8	1318	0.566	0.000	0.000	1	0.723	0.723
CD14	2028	0.241	0.000	0.000	0	0.389	0
DCs	372	0.261	0.000	0.000	0	0.412	0
pDC	180	−0.021	0.780	1.000	0	0.038	0
Platelets	33	0.165	0.358	1.000	0	0.262	0
NK_16hi	1649	0.062	0.011	0.000	0	0.117	0
NK_56hi	196	−0.031	0.658	1.000	0	0.058	0
B_cell	1358	0.067	0.013	0.000	0	0.126	0
gdT	1359	0.088	0.001	0.000	0	0.163	0
MAIT	285	0.127	0.031	1.000	0	0.225	0
CD16	648	−0.001	0.988	1.000	0	0.001	0
Treg	253	0.382	0.000	0.000	0	0.552	0
Lymph_prolif	61	0.089	0.492	1.000	0	0.152	0
HSC	26	NaN	NaN	NaN	NaN	NaN	NaN
Plasmablast	9	NaN	NaN	NaN	NaN	NaN	NaN
Mono_prolif	1	NaN	NaN	NaN	NaN	NaN	NaN

Table 2
GFPA analysis of EMT versus CD99 protein in CD4 T cells from 24 PBMC samples. GFPA score greater than 0.664430 and an adjusted p-value less than 0.01 were considered reliable.

Sample_id	Cell count	Correlation	P_value	Adj_p_value	Reliablity	GFPA_score	Adj_GFPA_score
MH8919282	2317	0.698	0.000	0.000	1	0.822	0.822
BGCV14_CV0940	729	0.635	9.55×10^{-84}	9.17×10^{-80}	1	0.777	0.777
newcastle65	2383	0.626	4.16×10^{-260}	4.00×10^{-256}	1	0.770	0.770
MH8919178	1426	0.620	2.96×10^{-152}	2.84×10^{-148}	1	0.765	0.765
MH8919332	2371	0.610	1.33×10^{-242}	1.28×10^{-238}	1	0.758	0.758
MH8919177	525	0.589	1.77×10^{-50}	1.70×10^{-46}	1	0.741	0.741
MH8919333	2305	0.569	1.76×10^{-198}	1.69×10^{-194}	1	0.725	0.725
BGCV05_CV0929	415	0.561	8.87×10^{-36}	8.52×10^{-32}	1	0.718	0.718
newcastle74	910	0.557	1.83×10^{-75}	1.76×10^{-71}	1	0.715	0.715
MH8919283	2054	0.547	4.56×10^{-161}	4.38×10^{-157}	1	0.707	0.707
MH8919179	848	0.522	1.20×10^{-60}	2.87×10^{-57}	1	0.686	0.686
BGCV04_CV0911	599	0.512	2.08×10^{-41}	3.99×10^{-38}	1	0.677	0.677
BGCV12_CV0926	1337	0.511	3.27×10^{-90}	1.04×10^{-86}	1	0.677	0.677
BGCV08_CV0915	920	0.506	3.09×10^{-61}	9.87×10^{-58}	1	0.672	0.672
MH8919176	921	0.504	1.49×10^{-60}	3.58×10^{-57}	1	0.670	0.670
BGCV01_CV0904	3473	0.500	1.80×10^{-219}	2.88×10^{-216}	1	0.666	0.666
BGCV15_CV0944	205	0.491	7.11×10^{-14}	1.71×10^{-10}	0	0.659	0
BGCV13_CV0934	1418	0.470	4.55×10^{-79}	1.09×10^{-75}	0	0.640	0
BGCV10_CV0939	746	0.436	4.18×10^{-36}	6.69×10^{-33}	0	0.608	0
BGCV09_CV0917	1456	0.418	5.56×10^{-63}	2.67×10^{-59}	0	0.590	0
BGCV02_CV0902	909	0.416	1.89×10^{-39}	1.82×10^{-36}	0	0.588	0
MH8919227	0	NaN	NaN	NaN	NaN	NaN	NaN
MH8919226	1	NaN	NaN	NaN	NaN	NaN	NaN
BGCV01_CV0902	0	NaN	NaN	NaN	NaN	NaN	NaN

GFPA.

GFPA has no restrictions on the type of proteins. In fact, any paired single-cell multi-omics data can be mined for associations using GFPA. However, non-paired data cannot be analyzed with GFPA. Association analysis of unpaired data across omics was a huge challenge. We plan to update the analysis of unpaired data in the next version. There is another limitation of GFPA. When researchers select gene collections with large amounts of data, it will cause GFPA to incur a large amount of runtime. The time overhead was mainly focused on the transformation of gene profiles into gene function centers and the calculation of GFPA scores.

In this study, we applied only Spearman correlation algorithm instead of Pearson correlation algorithm and Kandel correlation algorithm. The Pearson correlation algorithm requires the data to conform to a normal distribution, and the distribution of the gene profile and cell surface protein data does not satisfy this requirement. When researchers perform GFPA analysis using pre-processed data with a normal distribution, it is recommended to switch to the Spearman correlation algorithm. The target object of the Kandel correlation algorithm was ordered categorical variables. There was no ordering information in this study. When researchers pre-sort the data using the single-cell pseudo-time algorithm before executing GFPA, it is recommended to switch to the Kandel correlation algorithm. With these considerations in mind, we have retained the Pearson and Spearman correlation algorithms in the GFPA to help researchers perform a broader analysis.

In summary, an association analysis framework GFPA is proposed for single-cell multiomics data in this work. Considering that GFPA may consume a lot of computational resources with a large gene function set as the reference, we are optimizing the performance of GFPA to make it more efficient. We believe that GFPA will help researchers understand the function of cell surface proteins to reveal relevant disease progression.

Data availability

GFPA has been packaged as a python package, which provided cell subtype extraction, data preprocessing, gene database transformation, GFPA scoring, gene weight inference, visualization, and several other functions. Researchers can download and install GFPA from Pypi (http

s://pypi.org/project/gfpa/). In addition, GFPA was compatible with the python package "Scanpy" (1.9.1) [55]. The data processing process of GFPA was also similar to "Scanpy", and users can directly import scanpy objects (AnnData) to GFPA [55]. Further, we shared all the code of GFPA in our GitHub repository (https://github.com/studentiz/GFPA). All data can be accessed in figshare (https://doi.org/10.6084/m9.figshare.21901140). Anyone can use and modify GFPA for free (MIT license) [56]. Finally, we have published a tutorial for GFPA to help researchers quickly mine gene function and protein associations (https://github.com/studentiz/GFPA/tree/main/Tutorial).

Funding

This work was supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2021ZD0201900), the National Natural Science Foundation of China (Grant No. 12090052), Foundation of Education Department of Liaoning Province (Grant No. LJKZ0280), and the Fujian Province Foundation (Grant No. 2020Y4001).

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.106733.

References

[1] A.E. Murphy, N.G. Skene, A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis, Nat. Commun. 13 (2022) 7851.
[2] B. Owens, Genomics: the single life, Nature 491 (2012) 27–29.
[3] E. Mereu, A. Lafzi, C. Moutinho, et al., Benchmarking single-cell RNA-sequencing protocols for cell atlas projects, Nat. Biotechnol. 38 (2020) 747–755.
[4] C. Zhu, S. Preissl, B. Ren, Single-cell multimodal omics: the power of many, Nat. Methods 17 (2020) 11–14.
[5] C. Gawad, W. Koh, S.R. Quake, Single-cell genome sequencing: current state of the science, Nat. Rev. Genet. 17 (2016) 175–188.
[6] Y. Zhang, M.L. Amaral, C. Zhu, et al., Single-cell epigenome analysis reveals age-associated decay of heterochromatin domains in excitatory neurons in the mouse brain, Cell Res. 32 (2022) 1008–1021.
[7] R.T. Kelly, Single-cell proteomics: progress and prospects, Mol. Cell. Proteomics 19 (2020) 1739–1748.
[8] V. Marx, Method of the Year: spatially resolved transcriptomics, Nat. Methods 18 (2021) 9–14.
[9] Y. Hao, S. Hao, E. Andersen-Nissen, et al., Integrated analysis of multimodal single-cell data, Cell 184 (2021) 3573–3587.e3529.
[10] H. Hu, R. Liu, C. Zhao, et al., CITEMO(XMBD): a flexible single-cell multimodal omics analysis framework to reveal the heterogeneity of immune cells, RNA Biol. 19 (2022) 290–304.
[11] R. Lopez, J. Regier, M.B. Cole, et al., Deep generative modeling for single-cell transcriptomics, Nat. Methods 15 (2018) 1053–1058.
[12] H. Hu, Z. Feng, H. Lin, et al., Modeling and analyzing single-cell multimodal data with deep parametric inference, Briefings Bioinf. 24 (2023), bbad005.
[13] Z.-J. Cao, G. Gao, Multi-omics single-cell data integration and regulatory inference with graph-linked embedding, Nat. Biotechnol. 40 (2022) 1458–1466.
[14] B. Van de Sande, C. Flerin, K. Davie, et al., A scalable SCENIC workflow for single-cell gene regulatory network analysis, Nat. Protoc. 15 (2020) 2247–2276.
[15] C.B. Gonzalez-Blas, S. De Winter, G. Hulselmans, et al., SCENIC+: Single-Cell Multiomic Inference of Enhancers and Gene Regulatory Networks, bioRxiv, 2022, 504505, 2022.2008.2019.
[16] M. Stoeckius, C. Hafemeister, W. Stephenson, et al., Simultaneous epitope and transcriptome measurement in single cells, Nat. Methods 14 (2017) 865–868.
[17] V.M. Peterson, K.X. Zhang, N. Kumar, et al., Multiplexed quantification of proteins and transcripts in single cells, Nat. Biotechnol. 35 (2017) 936–939.
[18] E.P. Mimitou, A. Cheng, A. Montalbano, et al., Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells, Nat. Methods 16 (2019) 409–412.
[19] E.P. Mimitou, C.A. Lareau, K.Y. Chen, et al., Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells, Nat. Biotechnol. 39 (2021) 1246–1258.
[20] M. Stoeckius, S. Zheng, B. Houck-Loomis, et al., Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics, Genome Biol. 19 (2018) 224.
[21] E. Fiskin, C.A. Lareau, L.S. Ludwig, et al., Single-cell profiling of proteins and chromatin accessibility using PHAGE-ATAC, Nat. Biotechnol. 40 (2022) 374–381.
[22] Z. Hu, J. Yuan, M. Long, et al., The Cancer Surfaceome Atlas integrates genomic, functional and drug response data to identify actionable targets, Nat. Can. (Que.) 2 (2021) 1406–1422.
[23] G.M. Allen, W.A. Lim, Rethinking cancer targeting strategies in the era of smart cell therapeutics, Nat. Rev. Cancer 22 (2022) 693–702.
[24] F. Vergez, L. Largeaud, S. Bertoli, et al., Phenotypically-defined stages of leukemia arrest predict main driver mutations subgroups, and outcome in acute myeloid leukemia, Blood Cancer J. 12 (2022) 117.
[25] T. Wang, J. Sun, Q. Zhao, Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism, Comput. Biol. Med. 153 (2023), 106464.
[26] F. Sun, J. Sun, Q. Zhao, A deep learning method for predicting metabolite-disease associations via graph neural network, Briefings Bioinf. 23 (2022) bbac266.
[27] J. Hong, Y. Luo, Y. Zhang, et al., Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Briefings Bioinf. 21 (2020) 1437–1447.
[28] W.M. Wojtowicz, J. Vielmetter, R.A. Fernandes, et al., A human IgSF cell-surface interactome reveals a complex network of protein-protein interactions, Cell 182 (2020) 1027–1043, e1017.
[29] W. Wang, L. Zhang, J. Sun, et al., Predicting the potential human lncRNA-miRNA interactions based on graph convolution network with conditional random field, Briefings Bioinf. 23 (2022) bbac463.
[30] J. Fu, Q. Yang, Y. Luo, et al., Label-free proteome quantification and evaluation, Briefings Bioinf. 23 (2022) bbac477.
[31] E. Stephenson, G. Reynolds, R.A. Botting, et al., Single-cell multi-omics analysis of the immune response in COVID-19, Nat. Med. 27 (2021) 904–916.
[32] A. Liberzon, C. Birger, H. Thorvaldsdottir, et al., The Molecular Signatures Database (MSigDB) hallmark gene set collection, Cell Syst. 1 (2015) 417–425.
[33] A. Liberzon, A. Subramanian, R. Pinchback, et al., Molecular signatures database (MSigDB) 3.0, Bioinformatics 27 (2011) 1739–1740.
[34] B. Li, J. Tang, Q. Yang, et al., NOREVA: normalization and evaluation of MS-based metabolomics data, Nucleic Acids Res. 45 (2017) W162–w170.
[35] D. Thissen, L. Steinberg, D. Kuang, Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons, J. Educ. Behav. Stat. 27 (2002) 77–83.
[36] P. Qiu, Embracing the dropouts in single-cell RNA-seq analysis, Nat. Commun. 11 (2020) 1169.
[37] K.V. Wood, Marker proteins for gene expression, Curr. Opin. Biotechnol. 6 (1995) 50–58.
[38] A. Subramanian, P. Tamayo, V.K. Mootha, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 15545–15550.
[39] V.K. Mootha, C.M. Lindgren, K.-F. Eriksson, et al., PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, Nat. Genet. 34 (2003) 267–273.
[40] A. Kauffmann, T.F. Rayner, H. Parkinson, et al., Importing ArrayExpress datasets into R/bioconductor, Bioinformatics 25 (2009) 2092–2094.
[41] Q. Yang, B. Li, S. Chen, et al., MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis, J. Proteonomics 232 (2021), 104023.
[42] J. Tang, J. Fu, Y. Wang, et al., ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Briefings Bioinf. 21 (2020) 621–636.
[43] J. Lause, P. Berens, D. Kobak, Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data, Genome Biol. 22 (2021) 258.
[44] J. Vlassakis, L.L. Hansen, R. Higuchi-Sanabria, et al., Measuring expression heterogeneity of single-cell cytoskeletal protein complexes, Nat. Commun. 12 (2021) 4969.
[45] V. Hahaut, D. Pavlinic, W. Carbone, et al., Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq, Nat. Biotechnol. 40 (2022) 1447–1451.
[46] P. Pollard, J.T. Richardson, On the probability of making type I errors, Psychol. Bull. 102 (1987) 159–163.
[47] D.M. Dorfman, M. Kraus, A.R. Perez-Atayde, et al., CD99 (p30/32MIC2) immunoreactivity in the diagnosis of leukemia cutis, Mod. Pathol. 10 (1997) 283–288.
[48] L.A. Romero, T. Hattori, M.A.E. Ali, et al., High-valency anti-CD99 antibodies toward the treatment of T cell acute lymphoblastic leukemia, J. Mol. Biol. 434 (2022), 167402.
[49] J. Shi, Z. Zhang, H. Cen, et al., CAR T cells targeting CD99 as an approach to eradicate T-cell acute lymphoblastic leukemia without normal blood cells toxicity, J. Hematol. Oncol. 14 (2021) 162.
[50] P.M. Maciocia, P.A. Wawrzyniecka, B. Philip, et al., Targeting the T cell receptor β-chain constant region for immunotherapy of T cell malignancies, Nat. Med. 23 (2017) 1416–1423.
[51] B. Guo, H. Yan, L. Li, et al., Collagen triple helix repeat containing 1 (CTHRC1) activates Integrin β3/FAK signaling and promotes metastasis in ovarian cancer, J. Ovarian Res. 10 (2017) 69.
[52] C. Wang, Z. Li, F. Shao, et al., High expression of Collagen Triple Helix Repeat Containing 1 (CTHRC1) facilitates progression of oesophageal squamous cell

carcinoma through MAPK/MEK/ERK/FRA-1 activation, J. Exp. Clin. Cancer Res. 36 (2017) 84.
[53] J. Ye, W. Chen, Z.Y. Wu, et al., Upregulated CTHRC1 promotes human epithelial ovarian cancer invasion through activating EGFR signaling, Oncol. Rep. 36 (2016) 3588–3596.
[54] N. Sial, M. Ahmad, M.S. Hussain, et al., CTHRC1 expression is a novel shared diagnostic and prognostic biomarker of survival in six different human cancer subtypes, Sci. Rep. 11 (2021), 19873.
[55] F.A. Wolf, P. Angerer, F.J. Theis, Scanpy : large-scale single-cell gene expression data analysis, Genome Biol. 19 (2018) 1–5.
[56] J.H. Saltzer, D. Walden, The origin of the "MIT license, IEEE Ann. Hist. Comput. 42 (2020) 94–98.