



Article

An Efficient and Flexible Method for Deconvoluting Bulk RNA-Seq Data with Single-Cell RNA-Seq Data

Xifang Sun ¹, Shiquan Sun ^{2,3}  and Sheng Yang ^{4,*} 

¹ Department of Mathematics, School of Science, Xi'an Shiyou University, 710065 Xi'an, China; xfangsun@126.com

² School of Computer Science, Northwestern Polytechnical University, 710072 Xi'an, China; sqsun@nwpu.edu.cn

³ Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

⁴ Department of Biostatistics, School of Public Health, Nanjing Medical University, 211166 Nanjing, China

* Correspondence: yangsheng@njmu.edu.cn; Tel.: +86-2586868241

Received: 24 August 2019; Accepted: 26 September 2019; Published: 27 September 2019



Abstract: Estimating cell type compositions for complex diseases is an important step to investigate the cellular heterogeneity for understanding disease etiology and potentially facilitate early disease diagnosis and prevention. Here, we developed a computationally statistical method, referring to Multi-Omics Matrix Factorization (MOMF), to estimate the cell-type compositions of bulk RNA sequencing (RNA-seq) data by leveraging cell type-specific gene expression levels from single-cell RNA sequencing (scRNA-seq) data. MOMF not only directly models the count nature of gene expression data, but also effectively accounts for the uncertainty of cell type-specific mean gene expression levels. We demonstrate the benefits of MOMF through three real data applications, i.e., Glioblastomas (GBM), colorectal cancer (CRC) and type II diabetes (T2D) studies. MOMF is able to accurately estimate disease-related cell type proportions, i.e., oligodendrocyte progenitor cells and macrophage cells, which are strongly associated with the survival of GBM and CRC, respectively.

Keywords: cell-type compositions; deconvolution; single-cell RNA-seq; nonnegative matrix factorization; gene expression

1. Introduction

Accurate measurement of cell types in different tissues can often help our understanding of disease etiology and potentially facilitate the early diagnosis and prevention for complex diseases, especially for cancers [1–3]. For example, immune cells, such as CD8+ T cells, often proliferate in special tissues surrounding various types of tumors, mediating the immune response against tumor progression [4]. The traditional bulk samples are measured by the tissue-averaged gene expression levels, resulting in overlooked cellular heterogeneity [5]. The recent advance of single-cell RNA sequencing (scRNA-seq) technologies has allowed us to systemically characterize the heterogeneity of diverse cell types residing in tissues [6–8]. However, scRNA-seq studies are expensive and are only limited to a relatively small sample size [9], limiting the ability to investigate cellular heterogeneity across multiple individuals [10]. In contrast, bulk RNA-seq studies measure gene expression profiles for thousands of individuals [11–13]. Therefore, developing statistical methods for detecting cell type heterogeneity of existing large-scale bulk RNA-seq data with high-resolution scRNA-seq data plays an important role in understanding the interrelationship between cell type proportions and complex diseases [14].

To date, over sixty computational tools have been developed for deconvolution analysis, and these methods can be casted into two categories: reference-free and reference-based methods [15,16].

Reference-free methods dissect the heterogeneous samples into their constituent cell types with unsupervised schemes, i.e., without any prior reference knowledge [17,18]. For example, post-modified Negative Matrix Factorization (NMF) directly deconvolutes the gene expression levels of heterogeneous samples into the expected expression levels across the cell types and the corresponding cell type proportions using the alternating least square method [19]; Convex Analysis of Mixtures (CAM) identifies subpopulation marker genes from the original mixed gene expressions via convex analysis [20], etc. In contrast, reference-based methods estimate cell type proportions with supervised manner, i.e., with predefined cell type markers [21–23]. DeconRNASeq estimates mixed cell proportions with signature matrix through quadratic programming [24]; Cell type of Disease (CoD) quantifies disease-relevant immune cell compositions with a pre-defined list of 61 cell-surface markers based on random forest classification methods [25], etc.

Among these deconvolution methods, both MUlti-Subject SIngle Cell deconvolution (MuSiC) [10] and Cell-type Identification By Estimating Relative Subsets of RNA Transcripts (CIBERSORT) [26] would be able to directly model both bulk RNA-seq data and scRNA-seq data to estimate cell type proportions. However, both methods have several important limitations. First, they model normalized gene expression values and effectively treat observed sequencing data as a continuous outcome. However, the count nature of RNA-seq or scRNA-seq data display high mean-variance dependency [27,28]. Failing to account for the mean-variance dependency of RNA-seq data is known to lead to loss of power in sequencing data analysis [29,30]. Second, both methods treat the cell-type specific gene expression levels which are estimated from scRNA-seq data as reference gene signatures. However, bulk tissues display a heterogeneity of cell compositions across different individuals [10]. Failing to account for the heterogeneity of cell compositions across different individuals may lead to distort the underlying truth biological signals. Third, no tailored methods have been developed to jointly model both bulk RNA-seq data (mixture samples) and scRNA-seq data (cell type-specific expression). Existing approaches often treat calculating cell type-specific gene expression levels and estimating cell type compositions as two separate steps, despite the interconnection between these two different types of analyses [10,26]. Failing to jointly model both types of data may lead to the sub-optimal cell type proportion estimates. In addition, existing methods for signature matrix estimation relies on predefined cell markers or preselected differential expression genes, which may filter out informative genes, resulting in biased estimation of a signature matrix.

Here, we present a new computational tool, Multi-Omics Matrix Factorization (denoted as MOMF), to jointly model bulk RNA-seq data and scRNA-seq data to detect cell type compositions which potentially influence the survival processing effect. MOMF not only directly models raw gene expression counts of both bulk RNA-seq data and scRNA-seq data to avoid the biased cell composition estimations caused by normalization step, but also accounts for the heterogeneity of cell compositions across different individuals to estimate the underlying true cell compositions in bulk tissues. In addition, MOMF is not limited to bulk RNA-seq data and scRNA-seq data which are from the same experiment/study, e.g., bulk RNA-seq data is from The Cancer Genome Atlas (TCGA) database while scRNA-seq data can be from Gene Expression Omnibus (GEO) database. MOMF relies on a nonnegative matrix factorization (NMF) framework [31], using the alternating direction method of multipliers (ADMM) algorithm to infer the parameters in the model [32]. The overview and detailed algorithm of MOMF are shown in Results and Materials and Methods sections, respectively. Finally, we illustrate the benefits of MOMF with three in-depth analyses of real data applications, including the investigation of the relationship between cell compositions and survival statutes in glioblastoma (GBM), the relationship between cell compositions and survival statutes in colorectal cancer (CRC) and the relationship between cell compositions and Hb1Ac level in type II diabetes (T2D). From the results, we found that MOMF is able to accurately estimate two well-known cancer-related cell type proportions, i.e., oligodendrocyte progenitor cells (OPCs) and macrophage cells for the survival of GBM and CRC, respectively.

2. Methods and Materials

2.1. Model and Algorithm

Here, we jointly model both scRNA-seq data and bulk RNA-seq data to deconvolute mixed bulk RNA-seq data. The schematic view of the MOMF is shown in Figure 1. Specifically, we directly model both scRNA-seq count matrix X and bulk RNA-seq count matrix Y using Poisson distribution, as well as adding a prior distribution to account for the uncertainty of gene expression levels across different individuals in bulk RNA-seq data for parameter estimation. In particular, the gene expression count Y_{ij} for i 'th individual and j 'th gene in bulk RNA-seq data, we consider

$$Y_{ij} \sim \text{Poisson}(\mu_{ij}^y), i = 1, 2, \dots, n_y; j = 1, 2, \dots, p, \quad (1)$$

where Y_{ij} is the number of reads that measure the gene expression levels for j 'th gene and i 'th individual; n_y is the number of individuals; μ_{ij}^y is an unknown mean gene expression level for the i 'th individual and j 'th gene; and p is the number of genes; $\text{Poisson}(\cdot)$ represents the Poisson distribution.

The gene expression count X_{kj} for k 'th cell and j 'th gene in scRNA-seq data, we consider

$$X_{kj} \sim \text{Poisson}(\mu_{kj}^x), k = 1, 2, \dots, n_x; j = 1, 2, \dots, p, \quad (2)$$

where X_{kj} is the number of reads that measure the gene expression level for j 'th gene and k 'th cell; n_x is the number of cells; μ_{kj}^x is an unknown Poisson rate parameter that represents the underlying gene expression level for the k 'th cell and j 'th gene; and p is the number of genes; $\text{Poisson}(\cdot)$ represents the Poisson distribution.

In above models, we further decompose the unknown parameters μ_{ij}^y and μ_{kj}^x into two low-rank matrices, i.e.,

$$\mu_{ij}^y = \sum_{c=1}^C \Psi_{ic} W_{cj} + E_{ij}^y, i = 1, 2, \dots, n_y; j = 1, 2, \dots, p, \quad (3)$$

where Ψ_{ic} is the cell type-specific proportion for the i 'th individual and c 'th cell type; C is the number of cell types.

$$\mu_{kj}^x = \sum_{c=1}^C \Lambda_{kc} W_{cj} + E_{kj}^x, k = 1, 2, \dots, n_x; j = 1, 2, \dots, p, \quad (4)$$

where Λ_{kc} is the low-dimension structure for the k 'th cell and c 'th cell type; C is the number of cell type; the parameter W_{cj} is the element in the factor loading matrix that represents the underlying true cell-type specific gene expression level; the factor loading matrix W is shared between bulk RNA-seq and scRNA-seq data, allowing us to jointly model both data types and bypassing the estimation uncertainty inevitably occur in previous deconvolution methods; E_{ij}^y and E_{kj}^x are the residual terms that account for over-dispersion commonly observed in sequencing studies for bulk RNA-seq data and scRNA-seq data, respectively.

To account for the uncertainty of gene expression levels W in estimation step, we first estimate a reference gene expression panel h_{cj} for each cell type, i.e.,

$$h_{cj} = \frac{\sum_{k \in \Omega_c} X_{kj}}{\sum_{k=1}^{n_x} X_{kj}}, c = 1, 2, \dots, C; j = 1, 2, \dots, p, \quad (5)$$

where Ω_c is a set of the cells that belong to the cell type c . Then, we modeled the underlying true cell-type specific mean gene expression levels as

$$W_{cj} \sim TN(h_{cj}, \sigma^2), c = 1, 2, \dots, C; j = 1, 2, \dots, p, \quad (6)$$

where $TN(\cdot, \cdot)$ denotes the truncated normal distribution to guarantee that the cell type proportions are the non-negative values; the parameter σ^2 is an overall fixed parameter which is estimated from real

data to measure the uncertainty. In above model, we are interested in estimating the parameter Ψ_{ic} from bulk RNA-seq data for downstream analyses. The task requires the development of computational algorithms to infer the parameters. To reduce the computational burden of estimation, we used the Alternating Direction Method of Multipliers (ADMM) algorithm which has been widely applied for nonnegative matrix factorization problems [30] to infer the parameters.

To utilize the ADMM algorithm, we first construct the objective function

$$\begin{aligned} \mathcal{L} = & D(\mathbf{Y}|\mu^y) + \text{Tr}(\mathbf{U}^y(\mu^y - \Psi\mathbf{W})^T) + \frac{\rho}{2}\|\mu^y - \Psi\mathbf{W}\|_F^2 + \text{Tr}(\mathbf{U}^\Psi(\Psi - \Psi_+)^T) + \\ & \frac{\rho}{2}\|\Psi - \Psi_+\|_F^2 + D(\mathbf{X}|\mu^x) + \text{Tr}(\mathbf{U}^x(\mu^x - \Lambda\mathbf{W})^T) + \frac{\rho}{2}\|\mu^x - \Lambda\mathbf{W}\|_F^2 + \\ & \text{Tr}(\mathbf{U}^\Lambda(\Lambda - \Lambda_+)^T) + \frac{\rho}{2}\|\Lambda - \Lambda_+\|_F^2 + \text{Tr}(\mathbf{U}^W(\mathbf{W} - \mathbf{H})^T) + \frac{\rho}{2}\|\mathbf{W} - \mathbf{H}\|_F^2, \end{aligned} \tag{7}$$

where $D(y|x) = y\log(\frac{y}{x}) - y + x$ is the Kullback-Leibler (KL) divergence; $\mathbf{U}^y, \mathbf{U}^x, \mathbf{U}^\Psi, \mathbf{U}^\Lambda$ and \mathbf{U}^W are element-wise coefficients; Ψ_+ and Λ_+ are the non-negative matrix for Ψ and Λ , respectively; ρ is the penalty parameter; \mathbf{H} is reference gene expression panel; \mathbf{W} is underlying true gene expression panel; $\text{Tr}(\cdot)$ denotes the trace of a matrix. The updating equations for the parameters are as follows:

Taking the derivative of \mathcal{L} with respect to μ_{ij}^y and μ_{kj}^x , we have

$$\begin{cases} \mu_{ij}^y = \frac{\rho\Psi_{ic}W_{cj} - \mathbf{U}_{ij}^y - 1 + \sqrt{(\rho\Psi_{ic}W_{cj} - \mathbf{U}_{ij}^y - 1)^2 + 4\rho Y_{ij}}}{2\rho} \\ \mu_{kj}^x = \frac{\rho\Lambda_{kc}W_{cj} - \mathbf{U}_{kj}^x - 1 + \sqrt{(\rho\Lambda_{kc}W_{cj} - \mathbf{U}_{kj}^x - 1)^2 + 4\rho X_{kj}}}{2\rho} \end{cases} \tag{8}$$

Taking the derivative of \mathcal{L} with respect to Ψ_{ic} and Λ_{kc} , we have

$$\begin{cases} \Psi = (\mathbf{W}\mathbf{W}^T + \mathbf{I})^{-1}(\mathbf{Y}\mathbf{W}^T + \Psi_+ + \frac{1}{\rho}(\mathbf{U}^y\mathbf{W}^T - \mathbf{U}^\Psi)) \\ \Lambda = (\mathbf{W}\mathbf{W}^T + \mathbf{I})^{-1}(\mathbf{X}\mathbf{W}^T + \Lambda_+ + \frac{1}{\rho}(\mathbf{U}^x\mathbf{W}^T - \mathbf{U}^\Lambda)) \end{cases} \tag{9}$$

Taking the derivative of \mathcal{L} with respect to \mathbf{W} , we have

$$\mathbf{W} = (\Psi^T\Psi + \Lambda^T\Lambda + \mathbf{I})^{-1}\left(\Psi^T\mathbf{Y} + \Lambda^T\mathbf{X} + \mathbf{H} + \frac{1}{\rho}(\Psi^T\mathbf{U}^y + \Lambda^T\mathbf{U}^x - \mathbf{U}^W)\right) \tag{10}$$

Updating Ψ_+ , and Λ_+ with

$$\Psi_+ = \max\left(\Psi + \frac{1}{\rho}\mathbf{U}^y, 0\right), \Lambda_+ = \max\left(\Lambda + \frac{1}{\rho}\mathbf{U}^x, 0\right) \tag{11}$$

Updating the coefficients $\mathbf{U}^y, \mathbf{U}^x$, and \mathbf{U}^W with

$$\mathbf{U}^y \leftarrow \mathbf{U}^y + \rho(\mu^y - \Psi\mathbf{W}), \mathbf{U}^x \leftarrow \mathbf{U}^x + \rho(\mu^x - \Lambda\mathbf{W}), \mathbf{U}^W \leftarrow \mathbf{U}^W + \rho(\mathbf{W} - \mathbf{H}) \tag{12}$$

2.2. Simulation Designs

We performed benchmark experiments to examine the performance of MOMF and compared it with existing approaches, MuSiC and CIBERSORT. The cell type proportion matrix Ψ and the low-dimensional embedding matrix Λ were estimated from CRC data, including 590 individuals of bulk RNA-seq data and 359 cells of scRNA-seq data (details of CRC data in Methods and Materials). Following the model assumption, we first computed the expected gene expression levels of bulk RNA-seq data $\mathbb{E}(\mathbf{Y}) = \Psi\mathbf{W}$ and the expected gene expression levels of scRNA-seq data $\mathbb{E}(\mathbf{X}) = \Lambda\mathbf{W}$, where \mathbf{W} was randomly generated from gamma distribution with shape parameter 2 and inverse scale parameter 2 (i.e., R function *rgamma*). Then, we randomly generated \mathbf{Y} and \mathbf{X} from Poisson distribution (i.e., R function *rpois*). We simulated 10,000 genes and varied the number of cell types C to be either 2

(Epithelial and Macrophage), 3 (B cell, T cell and macrophage) and 5 (B cell, T cell, Epithelial, Fibroblast, Macrophage) to examine the performance of different deconvolution methods. Finally, we utilized Pearson correlation and mean of difference (MSE) between the estimated proportion \hat{p} to the ground truth p to measure the performance of different methods.

2.3. Bulk RNA-Seq and scRNA-Seq Data for GBM

Bulk RNA-seq data of GBM were downloaded from TCGA, which were measured on 56,716 transcripts and 153 individuals. We used the Level-3 Illumina Hiseq data which have performed quality control by TCGA workgroup. All data portals were accessed on March 2016. We filtered out the samples that do not include the survival information or survival time that equals to zero. scRNA-seq data (GSE67835) from brain tissue, consist of 466 cells, including astrocytes (62 cells), endothelial (20 cells), fetal quiescent (110 cells), fetal replicating (25 cells), hybrid (46 cells), microglia (16 cells), neurons (131 cells), oligodendrocytes (38 cells) and OPC (18 cells). We selected the transcripts that are larger than 5 and at least ten cells are expressed for each gene. Finally, we analyzed 144 individuals and 285 cells with common shared 11,120 genes for both bulk RNA-seq data and scRNA-seq data, respectively.

2.4. Bulk RNA-Seq and scRNA-Seq Data for CRC

Bulk RNA-seq data of CRC were downloaded from TCGA, measured on 56,716 transcripts and 616 individuals, including 453 Colon Adenocarcinoma (COAD) patients and 163 Rectum Adenocarcinoma (READ) patients. We filtered out the samples that do not include the survival information or survival time that equals to zero. Clinical information of both COAD and READ groups are described in Table 1. We used the Level-3 Illumina Hiseq data which were performed with quality control that was done by the TCGA workgroup. All the data portals were accessed on March 2016. Continuous variables were summarized as mean \pm standard deviation (SD), and categorized variables were described by frequency (n) and proportion (%). Endpoint is regarded as the death in the survival analysis. We used the log-rank test to compare the survival time of both COAD and READ groups.

scRNA-seq data (GSE81861) from CRC consist of 364 cells, including epithelial (272 cells), fibroblasts (17 cells), endothelial (4 cells), B (17 cells), T (34 cells), mast (1 cell) and macrophage (19 cells) [33]. Finally, we analyzed 590 individuals and 359 cells with common shared 33,888 transcripts for both bulk RNA-seq data and scRNA-seq data, respectively.

Table 1. Demographic distribution of discovery and validation study populations.

Variables	COAD (N = 435)	READ (N = 155)	<i>p</i>
Age (years), mean \pm SD	67.30 \pm 12.97	65.33 \pm 11.49	0.089
Gender, n (%)			0.680
Female	202 (46.43)	69 (44.52)	
Male	233 (53.56)	86 (55.48)	
Tumor stage (%)			0.166
0-I	240 (55.17)	76 (49.68)	
II-IV	184 (42.30)	71 (45.81)	
Unknown	11 (2.53)	8 (5.16)	
Race (%)			7.25×10^{-4}
White	207 (45.59)	77 (47.59)	
Non-white	70 (41.08)	6 (6.5)	
Unknown	158 (13.33)	71 (45.81)	
Survival year (month)			
Median	2532	1741	0.3
Dead, n (%)	97 (22.23)	25 (16.13)	

2.5. Bulk RNA-Seq and scRNA-Seq Data for T2D

We directly downloaded both processed T2D bulk RNA-seq and scRNA-seq data from MuSiC study [10]. For bulk RNA-seq data, we filtered out bulk individuals that do not have Hb1Ac level information. Based on clinical standard, Hb1Ac levels less than 6.0% is classified as normal sample while larger than 6.5% is classified as diabetic sample. For scRNA-seq data, we used five cell types, including the beta cell (171 cells), the delta cell (59 cells), the gamma cell (75 cells), and the ductal cell (135 cells), for downstream deconvolution analysis. Finally, we analyzed 77 individuals and 883 cells with common shared 14,934 transcripts for both bulk RNA-seq data and scRNA-seq data, respectively.

2.6. Software for Analyses

All calculations for the pooling of the effect estimates were performed using R program (version 3.6.1). TCGA data were downloaded by *TCGAbiolinks* package (version 2.13.3). The *k*-means clustering algorithm was performed on inferred cell type proportions using R function *kmeans* (iter.max = 10,000). Log-rank tests were performed by *survival* package (version 2.43.1). KM method was performed by *survminer* package (version 0.4.4) to illustrate the survival curves of different clusters. We used *biomaRt* package (version 2.40.1) to transfer Ensembl to gene symbol. MuSiC was performed by *MuSiC* R package (version 0.1.0) with default parameter settings. CIBERSORT was performed by the webserver tool (<https://cibersort.stanford.edu/>). MOMF requires the raw count gene expression matrices from both bulk RNA-seq and scRNA-seq studies, the penalty parameter ρ of ADMM algorithm was 2 and the number of iterations was 5000. Our method was implemented by the Rcpp as an R package. The source code of all experiments is freely available on GitHub <https://github.com/sqsun/MOMF>.

3. Results

3.1. Method Overview

We present a new computational method, MOMF, to estimate the cell type proportions across multiple individuals. An overview of MOMF is provided in Materials and Methods, and the details of the ADMM algorithm are provided in Supplementary Text. We also illustrate the schematic of the MOMF in Figure 1. Briefly, MOMF jointly models both bulk RNA-seq count matrix \mathbf{Y} and scRNA-seq count matrix \mathbf{X} to infer the cell compositions $\mathbf{\Psi}$ of bulk individuals and low-rank matrix $\mathbf{\Lambda}$ of scRNA-seq data via matrix factorization, i.e., $\mathbf{Y} = \mathbf{\Psi}\mathbf{W} + \mathbf{E}^y$ and $\mathbf{X} = \mathbf{\Lambda}\mathbf{W} + \mathbf{E}^x$, where \mathbf{E}^y and \mathbf{E}^x represent the residual errors for bulk RNA-seq data and scRNA-seq data, respectively. The common shared gene specific expression matrix \mathbf{W} between both bulk RNA-seq data and scRNA-seq data will be inferred from a reference signature expression level to accounting for the heterogeneity of cell compositions across different individuals. Our model starts with scRNA-seq data measured by a few hundreds of thousands cells and assumes that the cell type labels for the cells are known. MOMF deconvolutes the mixture bulk RNA-seq individuals with cell type-specific expression levels to obtain the proportions of the cell types in each individual.

Here, MOMF is able to overcome three drawbacks that MuSiC and CIBERSORT have: (1) it directly models mean-variance dependence of raw gene expression counts, avoiding to introduce systematic errors that may lead to spurious biological signals in downstream analysis; (2) it indirectly models the underlying true signature gene expression levels that follows mean cell type-specific expression levels to account for the heterogeneity of cell compositions across different individuals (i.e., Equation (6) in Methods and Materials), and variance σ^2 in Equation (6) is to account for the degree of uncertainty; (3) it jointly models the bulk RNA-seq data and scRNA-seq data to avoid the sub-optimal cell type proportion estimates (Methods and Materials). To demonstrate the stability of MOMF, we performed MOMF on the same data. We found that MOMF displays high correlation with two independent runs (i.e., $R^2 = 0.99$) (Supplementary Figure S1). Overall, MOMF is a more flexible model to estimate the cell type proportions of bulk RNA-seq data.

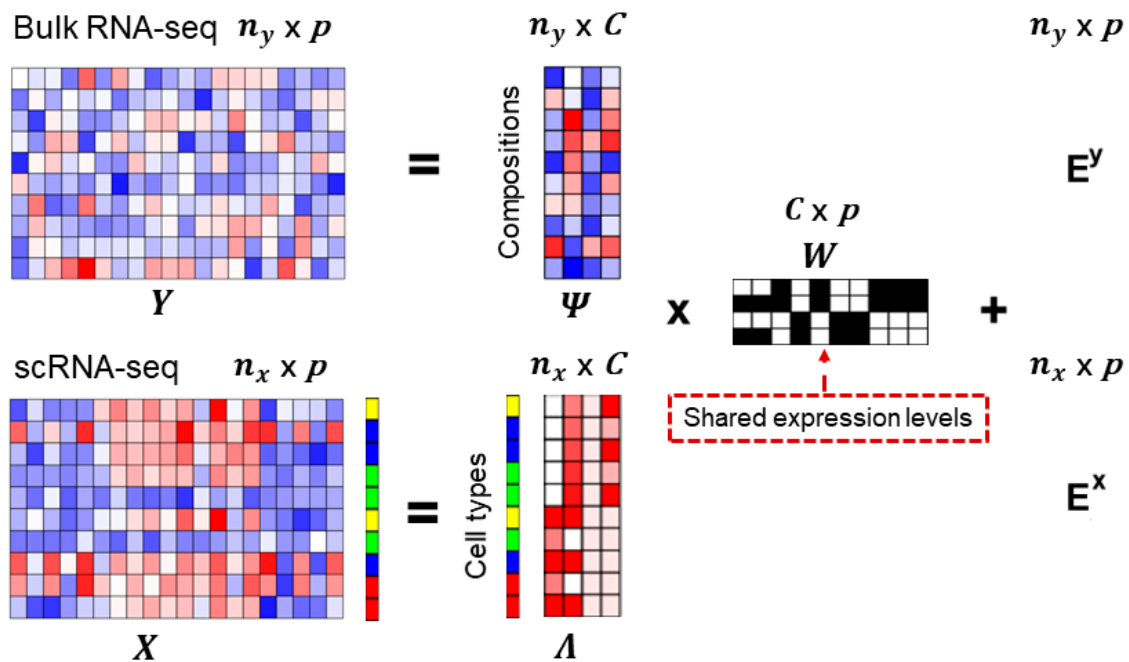


Figure 1. Overview of Multi-Omics Matrix Factorization (MOMF) framework. MOMF integrates bulk RNA-seq data and scRNA-seq data, to deconvolute the two expression matrices by the shared information and estimate the cell-type proportions for each individual. Specifically, MOMF jointly models both bulk RNA-seq count matrix Y and scRNA-seq count matrix X to infer the cell compositions Ψ of bulk RNA-seq data and low-rank matrix Λ of scRNA-seq data via matrix factorization, i.e., $Y = \Psi W + E^y$ and $X = \Lambda W + E^x$, where W is the common shared gene expression levels and E^y and E^x represent the residual errors for bulk RNA-seq data and scRNA-seq data, respectively. The heatmaps are used to illustrate the gene expression level (Y and X); cell specific expression levels (bulk RNA-seq: Ψ ; scRNA-seq: Λ); and gene specific expression levels (W). The color bar along with the heatmaps of scRNA-seq data represents the cell types. n_y is the number of individuals; n_x is the number of cells; p is the number of common shared genes; C is the number of cell types.

3.2. Normalization Distorts Raw Expression Counts

As we mentioned in the benefits of MOMF, proper normalization is a critical step that affects the estimation of the cell type proportions in deconvolution analysis. In bulk RNA-seq analysis, normalization has been extensively investigated [28,30,34]. In scRNA-seq data analysis, normalization is still an intractable problem [35–37]. The most popular normalization of scRNA-seq data is the log transformation of count per million (CPM) [38], i.e., $\log_2(y_{ij}/N_j + c)$ where y_{ij} is the gene expression level for i 'th cell and j 'th gene, N_j is the read sequencing depth, and c is a pseudocount. The logarithm transformation of CPM (logCPM) may distort the raw gene expression counts and introduces the zero-inflation artifacts due to the large number of zero counts [39]. For example, all the zeros remain $\log_2(1 + 0) = 0$, but the ones turn into value $\log_2(1 + 1/3000 \times 10^6) = \log_2(334) \approx 8.4$, and the counts that 10 will have value $\log_2(3340) \approx 11.7$. The large, artificial gap between zero and nonzero values makes the log-normalized data appear zero-inflated. To illustrate this phenomenon, we examined the distribution of an example gene (ENSG00000180725) in CRC scRNA-seq data before and after the logarithm transformation with varying normalizations (Supplementary Figure S2).

3.3. Simulations

We first evaluated the performance of different deconvolution methods on simulation studies with ground truth. To do so, we applied three methods, MOMF, MuSiC, and CIBERSORT to simulated datasets (Methods and Materials) and evaluated the performance of different methods based on the Pearson correlation and difference between estimated cell type proportion and ground truth. In the

analysis, we varied the number of cell types to be either 2, 3 or 5 to examine their influence on the accuracy of cell compositions. The evaluation results are summarized in Figure 2, Supplementary Figure S3 and Supplementary Figure S4.

From the results, we found that MOMF achieves the best performance across all parameter settings. For example, with the simulated data based on three cell types (B cells, T cells and Macrophage cells), the Pearson correlation R generated by MOMF is 0.992 and MSE is 0.040, while MuSiC (R is -0.753 and MSE is 0.611) and CIBERSORT (R is -0.213 and MSE is 0.334) do not fare well (Figure 2). With the small number of cell types (i.e., 2), all three methods are able to generate high Pearson correlations (Supplementary Figure S3). However, the scatter plots show the vertical patterns due to many zeros or ones and proportions were estimated by MuSiC and CIBERSORT. The Pearson correlation will decrease when increasing the number of cell types (i.e., 5). For example, the Pearson correlation R generated by MOMF is 0.618 and MSE is 0.135, while MuSiC (R is -0.368 and MSE is 0.490) and CIBERSORT (R is -0.692 and MSE is 0.395) do not fare well (Supplementary Figure S4).

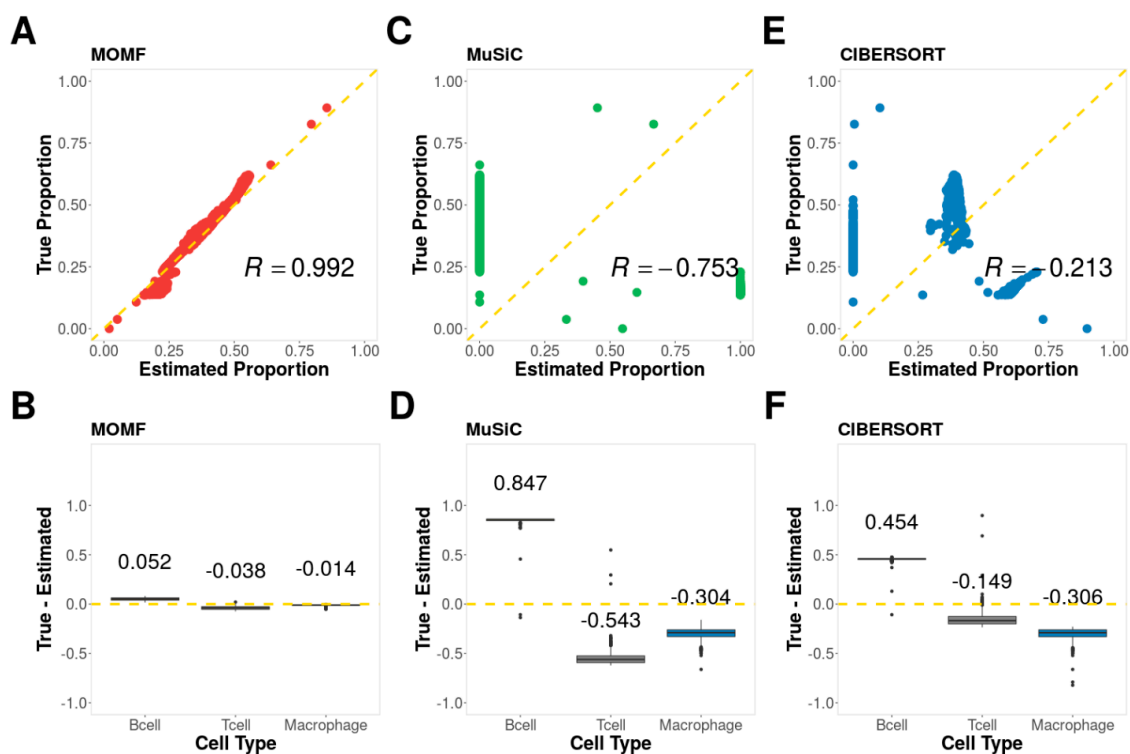


Figure 2. Simulation results. The simulated data based three cell types, B cells, T cells, and Macrophage cells. (A) The scatter plot of ground truth and cell type proportion estimated by MOMF; (B) The boxplot to show the difference between ground truth and cell type proportion estimated by MOMF (C) The scatter plot of ground truth and cell type proportion estimated by MuSiC; (D) The boxplot to show the difference between ground truth and cell type proportion estimated by MuSiC (E) The scatter plot of ground truth and cell type proportion estimated by CIBERSORT; (F) The boxplot to show the difference between ground truth and cell type proportion estimated by CIBERSORT. R : Pearson correlation.

3.4. Human Glioblastoma (GBM) Data

GBM is the most common primary brain tumor in adults [40–42]. GBM is characterized by the presence of hyperplastic blood vessels and the presence of small areas of necrotizing tissue that are surrounded by anaplastic cells [43]. Therefore, our primary goal here is to characterize how cell type proportions influence the survival time of GBM. We first applied MOMF on bulk RNA-seq data from GBM, which consist of 153 individuals and 60,486 transcripts and scRNA-seq data from healthy human brains, which consist of nine cell types and 18,752 transcripts (details in Materials and Methods). These cell types include astrocytes (62 cells), endothelial (20 cells), fetal quiescent (110 cells),

fetal replicating (25 cells), hybrid (46 cells), microglia (16 cells), neurons (131 cells), oligodendrocytes (38 cells) and OPC (18 cells) [44,45]. Following the preprocessing (Materials and Methods), for bulk RNA-seq data, we finally performed the analyses on 144 individuals along with median survival time (MST) 333 days; for scRNA-seq data, we finally performed the analyses on 285 cells from six different cell types, including astrocytes (62 cells), endothelial (20 cells), microglia (16 cells), neurons (131 cells), oligodendrocytes (38 cells) and OPC (18 cells). Finally, 11,120 transcripts commonly shared between both bulk RNA-seq data and scRNA-seq data were performed in the experiments.

We applied three deconvolution methods, MOMF, MuSiC, and CIBERSORT on both datasets to estimate the cell compositions across all GBM individuals. Average cell type proportions of astrocytes, endothelial, microglia, neurons, oligodendrocytes and OPC estimated from MOMF are roughly 3.6%, 25.2%, 4.9%, 4.7%, 3.2% and 57.9%, respectively; from MuSiC are 20.3%, 32.3%, 14.6%, 28.3%, 4.5% and 0.0%, respectively; and from CIBERSORT are 31.5%, 27.4%, 12.8%, 6.6%, 8.8% and 12.4%, respectively (Figure 3A). From the results, we found that MOMF provided higher OPCs cell type proportion (57.9%) than MuSiC (0.0%) and CIBERSORT (12.4%). OPCs are highly associated with cellular differentiation for oncogenic transformation in RCAS/tv-a model [41] and can be broadly arrayed in an adult brain where they constitute the largest pool of dividing cells [46]. The second-high cell type proportion is from endothelial cells. The cell type proportions from three methods MOMF, MuSiC, and CIBERSORT are 25.2%, 32.3% and 27.4%, respectively. The connection between neural stem cells and the endothelial compartment plays an important role in GBM. A significant proportion of the vascular endothelium has a neoplastic origin with increasing endothelial cells [47,48]. The high cell type proportions of OPCs might not be necessary to show the high performance of MOMF due to the ground truth of cell type proportion is being unknown. To validate the cell type proportions estimated from three different methods, we performed the association between the cell type proportion and the survival time of its individuals as another criterion to measure the performance of different deconvolution methods. MOMF produced statistically significant Kaplan-Meier (KM) plot with four potential subtypes ($N_{CL1} = 57$, $N_{CL2} = 33$, $N_{CL3} = 44$, $N_{CL4} = 10$) of GBM cancer (p -value = 0.007, log-rank test) (Figure 3B). With short median survival time, the cluster CL4 identified by MOMF has poor-prognosis for subtype of GBM cancer. While the p -value provided by MuSiC and CIBERSORT are 0.065 ($N_{CL1} = 60$, $N_{CL2} = 34$, $N_{CL3} = 19$, $N_{CL4} = 31$) and 0.170 ($N_{CL1} = 52$, $N_{CL2} = 39$, $N_{CL3} = 26$, $N_{CL4} = 27$), respectively (Figure 3B), failing to identify the potential subtype of GBM cancer.

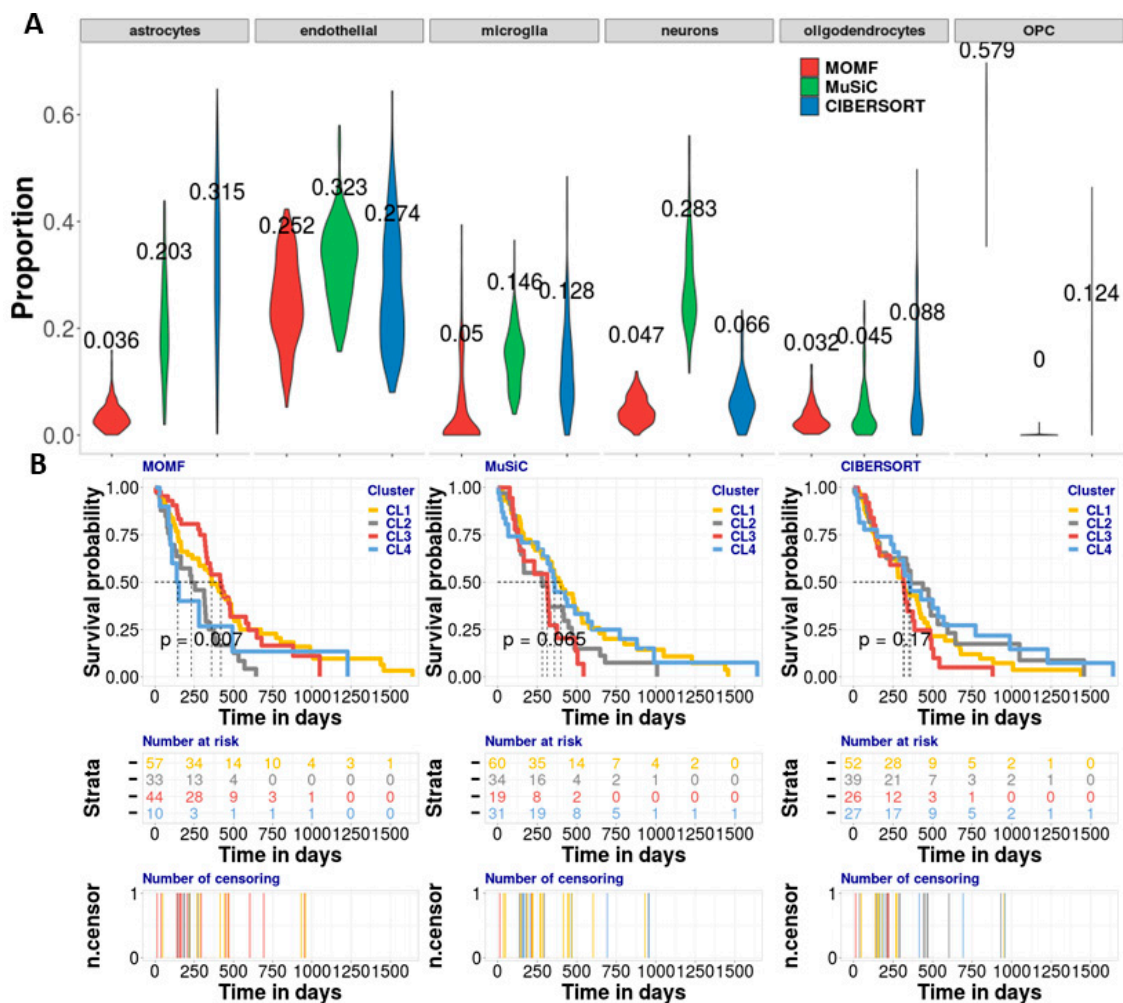


Figure 3. Analyzing GBM bulk RNA-seq data with brain scRNA-seq data. (A) The violin plot is to show the effect proportion of each cell type from three different deconvolution methods. We found that the OPCs and endothelial cells are enriched by MOMF, which means that the two cell types potentially contribute to the survival of GBM. (B) The KM plots are used to show the survival analysis for four clusters from the TCGA bulk RNAseq data. We use log-rank test to compare the distributions of four clusters. MOMF grouped the GBM samples into four subtypes (p -value = 0.007). The cluster CL4 is the poor-prognosis. Number at risk in the table shows the number of survival individuals at each 250 days. Number of censoring in the table shows censoring time of each individual. The numbers labeled in different colors in the two tables indicate the different subtypes.

3.5. Human Colorectal Cancer (CRC) Data

We next applied MOMF on CRC RNA-seq data, which consist of gene expression measurements from 56,716 transcripts and 616 individuals; scRNA-seq data which were measured on 364 cells from seven subpopulations (details in Materials and Methods). These seven subpopulations include T cells (34 cells), B cells (17 cells), epithelial cells (272 cells), fibroblast cells (17 cells), macrophage cells (19 cells), endothelial (4 cells), mast (1 cell) [33,49]. Following the preprocessing (Materials and Methods), for bulk RNA-seq data, we analyzed 590 individuals along with a median survival time (MST) of 2,532 days; for scRNA-seq data, we finally analyzed 359 cells of five cell types, T cells (34 cells), B cells (17 cells), epithelial cells (272 cells), fibroblast cells (17 cells) and macrophage cells (19 cells). Finally, we examined 33,888 common shared transcripts on both bulk RNA-seq data and scRNA-seq datasets.

To systematically benchmark the performance of MOMF, MuSiC and CIBERSORT, we applied them to CRC studies to estimate the cell-type proportion of bulk RNA-seq data across all individuals. We first

applied MOMF, MuSiC and CIBERSORT to estimate the cell type proportions of CRC bulk individuals (Figure 4A). From the results, we found that MOMF provides a higher macrophage cell proportion (14.9%), rather than MuSiC (2.0%) and CIBERSORT (5.1%). Tumor-associated macrophages (TAMs) are important components of the tumor microenvironment [50]. Macrophage cells may contribute to tumor growth and progression by promoting tumor cell proliferation and invasion, fostering tumor angiogenesis and suppressing antitumor immune cells [51]. The second high contribution of cell type proportion is the epithelial cell subpopulation (Figure 4A). The cell type proportions from MOMF, MuSiC and CIBERSORT are 49.7%, 90.0% and 73.6%, respectively. Small and large intestinal epithelium culture in vitro shows prolonged intestinal epithelial expansion with proliferation and multilineage differentiation [52]. The high cell type proportions of macrophage cells might be not necessary to show the high performance of MOMF due to the ground truth of cell type proportion is unknown. To validate the cell type proportions estimated from three different methods, we further performed the association analysis between the cell type proportion and its survival data, and the KM plot shows median survival time, number at risk and number of censoring of clustering results from the three methods. MOMF produced statistically significant KM plot with four potential subtypes ($N_{CL1} = 3$, $N_{CL2} = 90$, $N_{CL3} = 277$, $N_{CL4} = 220$) of CRC (p -value = 0.0013, log-rank test) (Figure 4B). With short median survival time, the cluster CL1 identified by MOMF has poor-prognosis for the subtype of CRC. While the p -value generated by MuSiC and CIBERSORT are 0.31 ($N_{CL1} = 23$, $N_{CL2} = 241$, $N_{CL3} = 228$, $N_{CL4} = 98$) and 0.098 ($N_{CL1} = 246$, $N_{CL2} = 113$, $N_{CL3} = 27$, $N_{CL4} = 204$), respectively (Figure 4B).

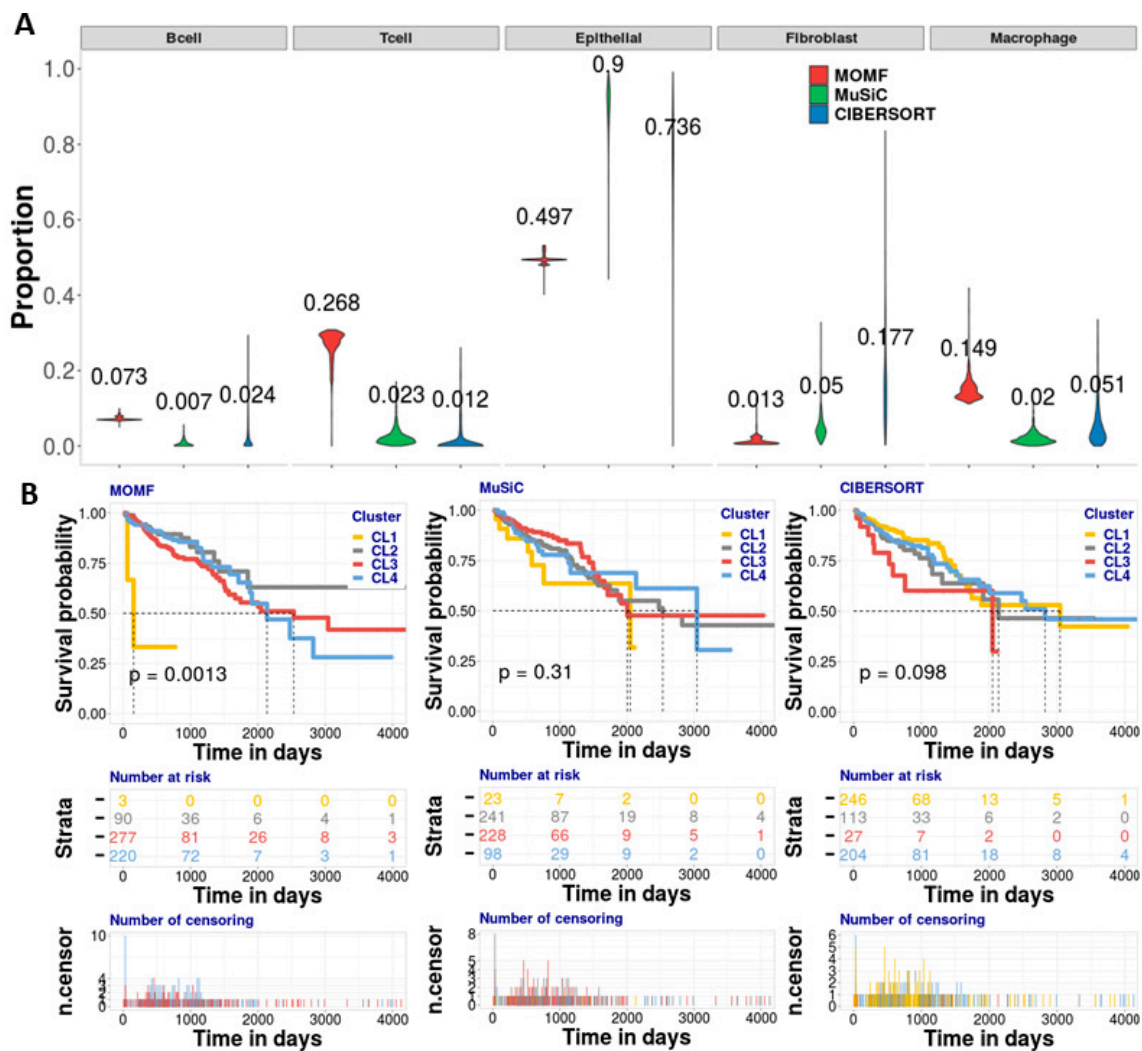


Figure 4. Analyzing CRC bulk RNA-seq data with colorectal cancer scRNA-seq data. (A) The violin plot is to show the effect proportion of each cell type from three different methods. We found that the epithelial, T and macrophage cells are enriched by MOMF, which means that the two cell types potentially contribute to the survival of CRC. (B) The KM plots are used to show the survival analysis for four clusters from the TCGA bulk data. We used log-rank test to compare the distributions of four clusters. MOMF grouped the CRC samples into four subtypes (p -value = 0.0013). The cluster CL1 shows the poor-prognosis. Number at risk in the table shows the number of survival individuals at each 1,000 days. Number of censoring in the table shows censoring time of each individual. The numbers labeled in different colors in the two tables indicate the different subtypes.

3.6. Human Type II Diabetes (T2D) Data

We finally applied MOMF on pancreatic islet studies: the bulk RNA-seq data consist of gene expression measurements of 32,581 transcripts and 89 individuals; the scRNA-seq data consist of 25,453 genes and 1097 cells from 14 subpopulations (details in Materials and Methods). The 14 subpopulations include alpha (443 cells), beta (171 cells), delta (59 cells), gamma (75 cells), ductal (135 cells), acinar (112 cells), co-expression (26 cells), endothelial (13 cells), epsilon (5 cells), mast (4 cells), MHC class II (1 cell), PSC (23 cells), unclassified (1 cell) and unclassified endocrine (29 cells) [33,49]. Following the preprocessing (Materials and Methods), for bulk RNA-seq data, we totally analyzed 77 individuals along with hemoglobin A1c (HbA1c) level; for scRNA-seq data, we analyzed 883 cells from five cell types, including 443 alpha cells, 171 beta cells, 59 delta cells, 75 gamma cells and 135 ductal cells. We examined 14,934 common shared genes on both bulk RNA-seq data and scRNA-seq data.

We first applied MOMF, MuSiC and CIBERSORT on pancreatic islets studies to recover the cell type proportion of bulk RNA-seq data across all individuals. Because bulk RNA-seq data contain T2D and control individuals, we examined the cell type proportions separately. From the results, we found that MOMF generated high ductal cell type proportions across T2D bulk individuals, and the percentage of ductal cell in T2D individuals is higher than that in normal individuals/controls (Figure 5A). Specifically, the recovered cell type proportions of MOMF, MuSiC and CIBERSORT are 97.4%, 53.1% and 35.8% in T2D individuals and 96.3%, 39.9% and 22.9% in controls, respectively. From the previous studies [53], our model validated the findings that the increased pancreatic ductal replication is strongly associated with T2D. The variance of cell type proportions estimated from MuSiC and CIBERSORT are extremely large. It is presumably due to both methods that are not able to account for the heterogeneity of cell compositions across different individuals.

To further validate the effectiveness of MOMF, MuSiC and CIBERSORT, we performed the association between cell type proportions and HbA1c levels. With cell type proportion results, we performed simple linear regression (*lm* function in R) with HbA1c levels, controlling for gender, age and body mass index (BMI) as the covariates. From the results, we found that the beta cell proportion shows the negative correlation with HbA1c levels while ductal cell proportion displays the positive correlation with HbA1c levels (Figure 5B). Both MOMF and MuSiC show the strong association with beta cell proportions, i.e., p -value = 0.004 and p -value = 0.006, respectively (Figure 5B), while CIBERSORT is weak, i.e., p -value = 0.683 (Figure 5B). A combination of increasing insulin resistance and reduced mass or dysfunction of the beta cells is the potential factor of T2D [54].

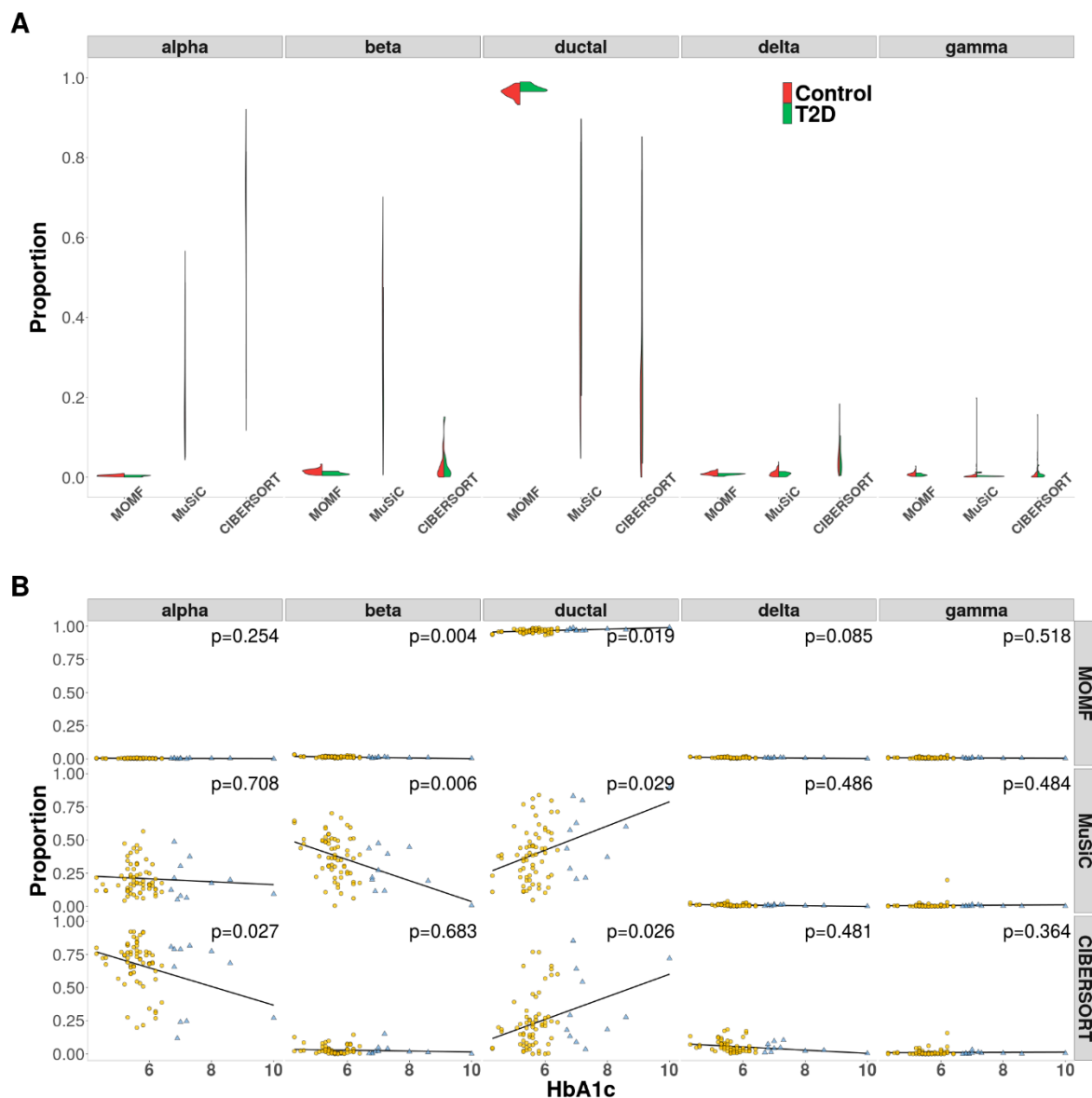


Figure 5. Analyzing T2D bulk RNA-seq data with pancreatic scRNA-seq data. **(A)** The violin plot is to show the effect proportion of each cell type from three different methods. beta and ductal cells are enriched from MOMF, which means that the two cell types potentially contribute to the survival of CRC. **(B)** The scatter plots are used to show the results of the associations between Hb1Ac level and cell proportion as adjust the covariates. The estimated beta cell proportions by both MOMF and MuSiC are strongly associated with Hb1Ac (p -value = 0.004 and 0.006).

4. Discussion

In this paper, we presented a new deconvolution method, MOMF, which directly models raw sequencing count data and accounts for the heterogeneity of cell compositions across different individuals. We have illustrated the benefits via performing deconvolution analysis through MOMF on both sequencing data (bulk RNA-seq and scRNA-seq). We have shown that MOMF is the only method currently available that can jointly model bulk RNA-seq and scRNA-seq data, providing the accurate measurement of cell type proportions which are highly associated with its corresponding survival times. With three in-depth analyses of real data applications, MOMF displays more reasonable and convincing results than existing two deconvolution methods, MuSiC and CIBERSORT. In contrast, MuSiC does not perform well on rare cell type proportion estimates, i.e., the inferred cell type proportions from

bulk RNA-seq are way off (very close to zero). Overall, MOMF is a useful and efficient tool, which is implemented as an R package for analyzing cell type compositions in tissue-specific gene expressions.

We primarily focused on using both RNA-seq data that are all modelling raw gene expression counts. One of the potential applications of MOMF is to deconvolute the mixed gene expression data which are from microarray experiments (i.e., continuous data). Therefore, exploring MOMF for continuous-counts or continuous-continuous scenarios is probably going to be part of our further work. In this case, MOMF will be useful to analyze the gene expression datasets. It can be easily extended to other multi-omics data analyses which can be either continuous data or count data [55,56]. This will probably be a very promising research direction in our further work.

MOMF is not without limitations. Perhaps the biggest limitation of MOMF is that the cell types are required to be labeled before applying MOMF. For scRNA-seq data, where there is no cell type label information, one solution is to first utilize the existing clustering methods, such as scNBMF or Seurat [57,58], to specify the cell type label for each cell, and then run MOMF with labeled scRNA-seq data.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2073-4409/8/10/1161/s1>, Supplementary Text: MOMF Deconvolution Method, Supplementary Figure S1: The correlation of cell type proportion estimated by MOMF with two independent runs on CRC data; Supplementary Figure S2: An example to show the distortion of normalized gene expression caused by logarithm transformation; Supplementary Figure S3: Simulation results with 2 cell types; Supplementary Figure S4: Simulation results with 5 cell types.

Author Contributions: S.S. conceived the study. S.Y. provided funding support. X.S., S.S. and S.Y. designed the experiments. X.S. and S.Y. adapted software, performed simulations and analyzed real data. X.S. and S.Y. wrote the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 81703321, 61902319), and the Fundamental Research Funds for the Central Universities (Grant No. 3102017OQD098), and the Natural Science Foundation of Shaanxi Province (Grant No. 2019JQ127).

Conflicts of 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.

Availability and Implementation: MOMF is implemented as an R package with source code freely available at: <https://github.com/sqsun/MOMF>.

References

1. Wagner, J.; Rapsomaniki, M.A.; Chevrier, S.; Anzeneder, T.; Langwieder, C.; Dykgers, A.; Rees, M.; Ramaswamy, A.; Muenst, S.; Soysal, S.D.; et al. A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* **2019**, *177*, 1330–1345.e1318. [[CrossRef](#)] [[PubMed](#)]
2. Van Hove, H.; Martens, L.; Scheyltjens, I.; De Vlaeminck, K.; Pombo Antunes, A.R.; De Prijck, S.; Vandamme, N.; De Schepper, S.; Van Isterdael, G.; Scott, C.L.; et al. A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment. *Nat. Neurosci.* **2019**, *22*, 1021–1035. [[CrossRef](#)] [[PubMed](#)]
3. Yuan, L.; Guo, F.; Wang, L.; Zou, Q. Prediction of tumor metastasis from sequencing data in the era of genome sequencing. *Brief. Funct. Genom.* **2019**. [[CrossRef](#)]
4. Smolders, J.; Heutinck, K.M.; Fransen, N.L.; Remmerswaal, E.B.; Hombrink, P.; Ten Berge, I.J.; van Lier, R.A.; Huitinga, I.; Hamann, J. Tissue-resident memory T cells populate the human brain. *Nat. Commun.* **2018**, *9*, 4593. [[CrossRef](#)] [[PubMed](#)]
5. Altschuler, S.J.; Wu, L.F. Cellular Heterogeneity: Do Differences Make a Difference? *Cell* **2010**, *141*, 559–563. [[CrossRef](#)] [[PubMed](#)]
6. Hou, Y.; Guo, H.; Cao, C.; Li, X.; Hu, B.; Zhu, P.; Wu, X.; Wen, L.; Tang, F.; Huang, Y. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **2016**, *26*, 304–319. [[CrossRef](#)] [[PubMed](#)]
7. Klein, A.M.; Mazutis, L.; Akartuna, I.; Tallapragada, N.; Veres, A.; Li, V.; Peshkin, L.; Weitz, D.A.; Kirschner, M.W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **2015**, *161*, 1187–1201. [[CrossRef](#)]

8. Hashimshony, T.; Senderovich, N.; Avital, G.; Klochendler, A.; de Leeuw, Y.; Anavy, L.; Gennert, D.; Li, S.; Livak, K.J.; Rozenblatt-Rosen, O. CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **2016**, *17*, 77. [[CrossRef](#)]
9. Ziegenhain, C.; Vieth, B.; Parekh, S.; Reinius, B.; Guillaumet-Adkins, A.; Smets, M.; Leonhardt, H.; Heyn, H.; Hellmann, I.; Enard, W. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **2017**, *65*, 631–643.e634. [[CrossRef](#)]
10. Wang, X.; Park, J.; Susztak, K.; Zhang, N.R.; Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **2019**, *10*, 380. [[CrossRef](#)]
11. Li, S.; Łabaj, P.P.; Zumbo, P.; Sykacek, P.; Shi, W.; Shi, L.; Phan, J.; Wu, P.-Y.; Wang, M.; Wang, C.; et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **2014**, *32*, 888–895. [[CrossRef](#)] [[PubMed](#)]
12. Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; Hasz, R.; Walters, G.; Garcia, F.; Young, N.; et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580–585. [[CrossRef](#)] [[PubMed](#)]
13. Taliun, D.; Harris, D.N.; Kessler, M.D.; Carlson, J.; Szpiech, Z.A.; Torres, R.; Taliun, S.A.G.; Corvelo, A.; Gogarten, S.M.; Kang, H.M.; et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* **2019**. [[CrossRef](#)]
14. Baron, M.; Veres, A.; Wolock, S.L.; Faust, A.L.; Gaujoux, R.; Vetere, A.; Ryu, J.H.; Wagner, B.K.; Shen-Orr, S.S.; Klein, A.M.; et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **2016**, *3*, 346–360.e344. [[CrossRef](#)] [[PubMed](#)]
15. Teschendorff, A.E.; Zheng, S.C. Cell-type deconvolution in epigenome-wide association studies: A review and recommendations. *Epigenomics* **2017**, *9*, 757–768. [[CrossRef](#)]
16. Mohammadi, S.; Zuckerman, N.; Goldsmith, A.; Grama, A. A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc. IEEE* **2016**, *105*, 340–366. [[CrossRef](#)]
17. Zuckerman, N.S.; Noam, Y.; Goldsmith, A.J.; Lee, P.P. A Self-Directed Method for Cell-Type Identification and Separation of Gene Expression Microarrays. *PLoS Comput. Biol.* **2013**, *9*, e1003189. [[CrossRef](#)]
18. Roy, S.; Lane, D.T.; Allen, C.; Aragon, A.D.; Werner-Washburne, M. A Hidden-State Markov Model for Cell Population Deconvolution. *J. Comput. Biol.* **2006**, *13*, 1749–1774. [[CrossRef](#)]
19. Liu, Y.; Liang, Y.; Kuang, Q.; Xie, F.; Hao, Y.; Wen, Z.; Li, M. Post-modified non-negative matrix factorization for deconvoluting the gene expression profiles of specific cell types from heterogeneous clinical samples based on RNA-sequencing data. *J. Chemom.* **2018**, *32*, e2929. [[CrossRef](#)]
20. Wang, N.; Hoffman, E.P.; Chen, L.; Chen, L.; Zhang, Z.; Liu, C.; Yu, G.; Herrington, D.M.; Clarke, R.; Wang, Y. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **2016**, *6*, 18909. [[CrossRef](#)]
21. Li, B.; Severson, E.; Pignon, J.-C.; Zhao, H.; Li, T.; Novak, J.; Jiang, P.; Shen, H.; Aster, J.C.; Rodig, S.; et al. Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.* **2016**, *17*, 174. [[CrossRef](#)] [[PubMed](#)]
22. Pollara, G.; Murray, M.J.; Heather, J.M.; Byng-Maddick, R.; Guppy, N.; Ellis, M.; Turner, C.T.; Chain, B.M.; Noursadeghi, M. Validation of Immune Cell Modules in Multicellular Transcriptomic Data. *PLoS ONE* **2017**, *12*, e0169271. [[CrossRef](#)] [[PubMed](#)]
23. Finotello, F.; Mayer, C.; Plattner, C.; Laschober, G.; Rieder, D.; Hackl, H.; Krogsdam, A.; Loncova, Z.; Posch, W.; Wilflingseder, D.; et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **2019**, *11*, 34. [[CrossRef](#)] [[PubMed](#)]
24. Gong, T.; Szustakowski, J.D. DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **2013**, *29*, 1083–1085. [[CrossRef](#)] [[PubMed](#)]
25. Frishberg, A.; Steuerman, Y.; Gat-Viks, I. CoD: Inferring immune-cell quantities related to disease states. *Bioinformatics* **2015**, *31*, 3961–3969. [[CrossRef](#)] [[PubMed](#)]
26. Newman, A.M.; Liu, C.L.; Green, M.R.; Gentles, A.J.; Feng, W.; Xu, Y.; Hoang, C.D.; Diehn, M.; Alizadeh, A.A. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **2015**, *12*, 453–457. [[CrossRef](#)]
27. Risso, D.; Perraudeau, F.; Gribkova, S.; Dudoit, S.; Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **2018**, *9*, 284. [[CrossRef](#)]
28. Sun, S.; Hood, M.; Scott, L.; Peng, Q.; Mukherjee, S.; Tung, J.; Zhou, X. Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res.* **2017**, *45*, e106. [[CrossRef](#)]

29. Amrhein, L.; Harsha, K.; Fuchs, C. A mechanistic model for the negative binomial distribution of single-cell mRNA counts. *bioRxiv* **2019**. [[CrossRef](#)]
30. Sun, S.; Zhu, J.; Mozaffari, S.; Ober, C.; Chen, M.; Zhou, X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. *Bioinformatics* **2019**, *35*, 487–496. [[CrossRef](#)]
31. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [[CrossRef](#)] [[PubMed](#)]
32. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **2011**, *3*, 1–122. [[CrossRef](#)]
33. Li, H.; Courtois, E.T.; Sengupta, D.; Tan, Y.; Chen, K.H.; Goh, J.J.L.; Kong, S.L.; Chua, C.; Hon, L.K.; Tan, W.S. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **2017**, *49*, 708–718. [[CrossRef](#)] [[PubMed](#)]
34. Maza, E.; Frasse, P.; Senin, P.; Bouzayen, M.; Zouine, M. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. *Commun. Integr. Biol.* **2013**, *6*, e25849. [[CrossRef](#)] [[PubMed](#)]
35. Cole, M.B.; Risso, D.; Wagner, A.; DeTomaso, D.; Ngai, J.; Purdom, E.; Dudoit, S.; Yosef, N. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Syst.* **2019**, *8*, 315–328. [[CrossRef](#)]
36. Ding, B.; Zheng, L.; Wang, W. Assessment of Single Cell RNA-Seq Normalization Methods. *G3 Genes Genomes Genet.* **2017**, *7*, 2039–2045. [[CrossRef](#)]
37. Vallejos, C.A.; Risso, D.; Scialdone, A.; Dudoit, S.; Marioni, J.C. Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Methods* **2017**, *14*, 565–571. [[CrossRef](#)]
38. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106. [[CrossRef](#)]
39. Townes, F.W.; Hicks, S.C.; Aryee, M.J.; Irizarry, R.A. Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model. *bioRxiv* **2019**. [[CrossRef](#)]
40. Llaguno, S.R.A.; Wang, Z.; Sun, D.; Chen, J.; Xu, J.; Kim, E.; Hatanpaa, K.J.; Raisanen, J.M.; Burns, D.K.; Johnson, J.E. Adult lineage-restricted CNS progenitors specify distinct glioblastoma subtypes. *Cancer Cell* **2015**, *28*, 429–440.
41. Lindberg, N.; Kastemar, M.; Olofsson, T.; Smits, A.; Uhrbom, L. Oligodendrocyte progenitor cells can act as cell of origin for experimental glioma. *Oncogene* **2009**, *28*, 2266–2275. [[CrossRef](#)] [[PubMed](#)]
42. Yuan, X.; Curtin, J.; Xiong, Y.; Liu, G.; Waschmann-Hogiu, S.; Farkas, D.L.; Black, K.L.; Yu, J.S. Isolation of cancer stem cells from adult glioblastoma multiforme. *Oncogene* **2004**, *23*, 9392–9400. [[CrossRef](#)] [[PubMed](#)]
43. Takano, S. Glioblastoma angiogenesis: VEGF resistance solutions and new strategies based on molecular mechanisms of tumor vessel formation. *Brain Tumor Pathol.* **2012**, *29*, 73–86. [[CrossRef](#)] [[PubMed](#)]
44. The Cancer Genome Atlas Research Network; McLendon, R.; Friedman, A.; Bigner, D.; Van Meir, E.G.; Brat, D.J.M.; Mastrogiannis, G.; Olson, J.J.; Mikkelsen, T.; Lehman, N.; et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **2008**, *455*, 1061–1068. [[CrossRef](#)]
45. Darmanis, S.; Sloan, S.A.; Zhang, Y.; Enge, M.; Caneda, C.; Shuer, L.M.; Gephart, M.G.H.; Barres, B.A.; Quake, S.R. A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 7285–7290. [[CrossRef](#)] [[PubMed](#)]
46. Zong, H.; Verhaak, R.G.W.; Canoll, P. The cellular origin for malignant glioma and prospects for clinical advancements. *Expert Rev. Mol. Diagn.* **2012**, *12*, 383–394. [[CrossRef](#)] [[PubMed](#)]
47. Ricci-Vitiani, L.; Pallini, R.; Biffoni, M.; Todaro, M.; Iavernici, G.; Cenci, T.; Maira, G.; Parati, E.A.; Stassi, G.; Larocca, L.M.; et al. Tumour vascularization via endothelial differentiation of glioblastoma stem-like cells. *Nature* **2010**, *468*, 824–828. [[CrossRef](#)] [[PubMed](#)]
48. Wang, R.; Chadalavada, K.; Wilshire, J.; Kowalik, U.; Hovinga, K.E.; Geber, A.; Fligelman, B.; Leversha, M.; Brennan, C.; Tabar, V. Glioblastoma stem-like cells give rise to tumour endothelium. *Nature* **2010**, *468*, 829–833. [[CrossRef](#)]
49. The Cancer Genome Atlas Network; Muzny, D.M.; Bainbridge, M.N.; Chang, K.; Dinh, H.H.; Drummond, J.A.; Fowler, G.; Kovar, C.L.; Lewis, L.R.; Morgan, M.B.; et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **2012**, *487*, 330–337. [[CrossRef](#)]

50. Zhang, R.; Qi, F.; Zhao, F.; Li, G.; Shao, S.; Zhang, X.; Yuan, L.; Feng, Y. Cancer-associated fibroblasts enhance tumor-associated macrophages enrichment and suppress NK cells function in colorectal cancer. *Cell Death Dis.* **2019**, *10*, 273. [[CrossRef](#)]
51. Engblom, C.; Pfirschke, C.; Pittet, M.J. The role of myeloid cells in cancer therapies. *Nat. Rev. Cancer* **2016**, *16*, 447–462. [[CrossRef](#)] [[PubMed](#)]
52. Ootani, A.; Li, X.; Sangiorgi, E.; Ho, Q.T.; Ueno, H.; Toda, S.; Sugihara, H.; Fujimoto, K.; Weissman, I.L.; Capecchi, M.R.; et al. Sustained in vitro intestinal epithelial culture within a Wnt-dependent stem cell niche. *Nat. Med.* **2009**, *15*, 701–706. [[CrossRef](#)] [[PubMed](#)]
53. Butler, A.; Galasso, R.; Matveyenko, A.; Rizza, R.; Dry, S.; Butler, P. Pancreatic duct replication is increased with obesity and type 2 diabetes in humans. *Diabetologia* **2010**, *53*, 21–26. [[CrossRef](#)] [[PubMed](#)]
54. Segerstolpe, Å.; Palasantza, A.; Eliasson, P.; Andersson, E.-M.; Andréasson, A.-C.; Sun, X.; Picelli, S.; Sabirsh, A.; Clausen, M.; Bjursell, M.K.; et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **2016**, *24*, 593–607. [[CrossRef](#)] [[PubMed](#)]
55. Sun, S.; Sun, X.; Zheng, Y. Higher-order partial least squares for predicting gene expression levels from chromatin states. *BMC Bioinform.* **2018**, *19*, 113. [[CrossRef](#)] [[PubMed](#)]
56. Jiang, J.; Xing, F.; Wang, C.; Zeng, X.; Zou, Q. Investigation and development of maize fused network analysis with multi-omics. *Plant Physiol. Biochem.* **2019**, *141*, 380–387. [[CrossRef](#)]
57. Sun, S.; Chen, Y.; Liu, Y.; Shang, X. A fast and efficient count-based matrix factorization method for detecting cell types from single-cell RNAseq data. *BMC Syst. Biol.* **2019**, *13*, 28. [[CrossRef](#)]
58. Butler, A.; Hoffman, P.; Smibert, P.; Papalexi, E.; Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **2018**, *36*, 411–420. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).