Summer 8-17-2023

A Review of Recent Gene Expression-Based and DNA Methylation-Based Mathematical Cell Type Deconvolution Methods

Chenxiao Tian

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics and Statistics

A Review of Recent Gene Expression-Based and DNA Methylation-Based Mathematical Cell Type Deconvolution Methods

By

Chenxiao Tian

A thesis presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Master of Arts

August 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

An acknowledgments page must be included in your final thesis. This page is where you should mention your thesis committee. If you wish to include a special dedication, you can either use it to close the acknowledgments page or place it on the page that immediately follows. The acknowledgments page should be listed in the table of contents. Place it after the final list used in the document, and before any dedication, abstract, or epigraph that is included.

It is appropriate to acknowledge sources of academic and financial support; some fellowships and grants require acknowledgment.

We offer special thanks to the Washington University School of Engineering for allowing us to use their dissertation and thesis template as a starting point for the development of this document.

Chenxiao Tian

Washington University in St. Louis

August 2023

Dedicated to my parents.

ABSTRACT OF THE THESIS

A Review of Recent Gene Expression-Based and DNA Methylation-Based Mathematical Cell Type Deconvolution Methods

by

Chenxiao Tian

Master of Arts in Mathematics and Statistics

Washington University in St. Louis, 2023

Professor Feres Renato, Chair

In recent years, many cell type deconvolution methods based on DNA methylation data and gene expression data have been developed. Both of these two methods have its special advantages and disadvantages, e.g., DNA methylation-based methods' data source is usually more stable than gene expression and DNA methylation is easier to measure in FFPE tissues or formalin-fixed paraffin-embedded, while some gene-expression data like scRNA-seq data usually has high cost and complexity. On the other hand, gene expression-based deconvolution methods currently have many more available methods than DNA methylation-based deconvolution methods, which leads to DNA methylation-based methods in many cases can learn from the existing gene expression-based methods, e.g., the EMeth learns from ICeD-T while the MethylCIBERSORT learns from CIBERSORT. Since both of these two kinds of different data-based methods are powerful tools to realize the purpose of cell type-specific deconvolution and may could benefit each other's development, as well as they have been still rapidly developing in recent years with believably more coming new methods in the future. It may be well worth looking back and comparing some recent gene expression data-based and DNA methylation-based deconvolution methods to get some comprehensive sense of this field's development and directions on both two different data-based deconvolution methods

A Review of Recent Gene Expression-Based and DNA Methylation-Based Mathematical Cell Type Deconvolution Methods

Chenxiao Tian[1]

Department of Mathematics, Washington University in St.Louis,

---

Abstract

In recent years, many cell type deconvolution methods based on DNA methylation data and gene expression data have been developed. Both of these two methods have its special advantages and disadvantages, e.g., DNA methylation-based methods' data source is usually more stable than gene expression and DNA methylation is easier to measure in FFPE tissues or formalin-fixed paraffin-embedded, while some gene-expression data like scRNA-seq data usually has high cost and complexity. On the other hand, gene expression-based deconvolution methods currently have many more available methods than DNA methylation-based deconvolution methods, which leads to DNA methylation-based methods in many cases can learn from the existing gene expression-based methods, e.g., the EMeth learns from ICeD-T while the MethylCIBERSORT learns from CIBERSORT. Since both of these two kinds of different data-based methods are powerful tools to realize the purpose of cell type-specific deconvolution and may could benefit each other's development, as well as they have been still rapidly developing in recent years with believably more coming new methods in the future. It may be well worth looking back and comparing some recent gene expression data-based and DNA methylation-based deconvolution methods to get some comprehensive sense of this field's development and directions on both two different data-based deconvolution methods.

Keywords: Gene Expression-Based, DNA Methylation-Based, ScRNA-seq, Bulk RNA-seq, Deconvolution Methods, Cell Type-specific Analysis, Cell Type-Specific Proportions, Cell Type-Specific Gene Expression, Cell Type-Specific Differential Expression Analysis

---

Background ＆Introduction

Since both the gene expression and the DNA methylation of cells vary across cell types, gene expression data and DNA methylation data can be both used for bulk cell

[1] **Email：chenxiao.t@wustl.edu**

samples analysis, like cell type proportions/compositions estimation or cell type-specific gene expression estimation and cell type-specific differential expression testing. In fact, DNA methylation data and gene expression data have following simple corresponding relationship in Tab.1. Such kinds of cell type specific analysis for bulk tissue samples are sometime very important for some other kinds of cell data analysis or more practical applications in biological problems.

Categories of the Regression Model Data	Response Variable	Dependent Variable	Covariate	Biology Basic
DNA Methylation Data-Based Regression Model	Cell Type-specific Methylation Expression	Observed Bulk Methylation Expression	Proportion of Special cell type	CpGs [2] of Cells
Gene Expression Data-Based Regression Model	Cell Type-specific Gene Expression	Observed Bulk Gene Expression	Proportion of Special cell type	Genes of Cells

Table 1: Comparisons of two different data based General regression model

For example, on the side of cell type proportions/compositions estimation analysis, if we obtain some reasonable and precise estimation of the cell type proportion in the bulk cell samples, then it would be very useful for some -omic data analysis or clinical studies. Like in the Epigenome Wide Association Studies, the consideration of cell type proportion variation is important, in [1.GB14], since blood can be viewed as a heterogeneous collection of different cell types and is exactly a kind of bulk cell sample, each with a very different DNA methylation profile. They examine data from some previous published research and find strong evidence of cell composition change across age in blood, so their findings underscore the importance of considering cell composition variability in epigenetic studies based on whole blood and other heterogeneous tissue sources.

For another example, on the other side of analyzing cell type-specific gene expression, for example, since unfortunately bulk samples usually contain many distinct cell types, if we want to identify genes with different expression levels between cancer samples versus controls, instead of considering the real bulk situation, we sometimes may only assume that the measured gene expression is just from the main cell type composition of the bulk tissue. In this case, the specific analysis of cell type-specific gene expression and the cell type proportion will both be very important for us to identify the genes with different expression levels in the real bulk situation like the cancer samples. Like in [2.FG20], the researchers use cell gene expression analysis to study the elucidation of mechanisms of topotecan-Induced Cell Death in human breast MCF-7 cancer cells. Their research identified several genes, FDXR, MSR, GSR, and GPx, which are involved in maintenance of cellular homeostasis due to increased ROS

[2] **CPG site refers to a region of DNA where the base sequence appears as cytosine followed by guanine. "CpG" is the abbreviation of "- C - phosphoric acid - G -".**

formation, were differentially expressed by TPT.

In this review, though we will be more focused on the methods' functions focused on the side of cell type composition estimation, i.e., the estimation of the cell type proportions/compositions. However, since the proportion of different cell type and the expression of the cell-type specific gene are highly connected and together to present the final bulk gene expression results in theory and also in assayed data by their matrix multiplication. So some methods have more fruitful functions, some methods aim at simultaneously estimating cell type-specific gene expression profiles and cell type proportions, even also can conduct the cell type-specific differential expression testing, e.g., the recent SCADIE method in [3.GB22].

Among those cell type composition estimation methods, some are based on the DNA methylation data while others are based on the gene expression data. Either of them has its special advantages, e.g., one advantage to estimate cell type proportion using DNA methylation rather than gene expression is that DNA methylation is usually more stable than gene expression and DNA methylation is easier to measure in FFPE tissues or formalin-fixed paraffin-embedded[4.BM10], which is the most commonly used form to store tissue samples and sometimes leads to a lower cost of DNA methylation compared to gene expression data, e.g., scRNA-seq data usually has high cost and complexity which leads to the necessity of getting the data just from bulk samples to reduce the cost, i.e., bulk RNA-seq data, this kind of data leads to some necessary extra deconvolutional analysis steps and methods. In contrast, gene expression data like scRNA-seq data or bulk RNA-seq data also has special advantages, for example, there are more mature computational deconvolution methods based on gene expression data which may also guide the deconvolution methods based on the DNA methylation data.

ScRNA-seq and bulk RNA-seq data in some cases can work together in a cell type deconvolution method and are not absolutely separated from each other, e.g., like the deconvolution methods CMP[11.NM19] and MuSiC [12.NC19], both of which are methods that utilize cell-type specific gene expression from single-cell RNA sequencing (scRNA-seq) data as references to characterize cell type compositions from bulk RNA-seq data in complex tissues, i.e., in some methods they take cell type-specific scRNA-seq as reference data to develop reference-based methods. While some other gene expression-based methods, for example DeCompress[28.NAR21]is a reference-free or semi-reference-free method. This differences between reference-based and reference-free are also common to see in DNA methylation-based deconvolution methods, e.g. EMeth is a reference-based methods and the methods in [16.Bio14], [17.GB18], [18.GB19], [27.GB19] are reference-free DNA methylation-based deconvolution methods.

What's more, beyond the gene expression-based deconvolution methods[5. FCDB20] and other traditional gene expression data applications mainly focused on the dissection of cell types/states, developmental trajectory, gene regulatory network, and alternative splicing. In recent years, there are more fruitful additional applications of scRNA-seq data and other gene expression data, such as Cell-to-cell communication network inference, Reconstruction of spatial cellular communications and gene expression, Identification of large-scale copy number variations, Analysis of single nucleotide

variants and RNA editing, Profiling long non-coding RNAs and circular RNAs… In words, comparing to DNA methylation, single-cell RNA-seq (scRNA-seq) technologies and related bioinformatics methods have been developing and innovating rapidly, which significantly revolutionized our understanding of the expression heterogeneity and transcriptome dynamics of individual cells gene, scRNA-seq data analysis currently is more mature and have more profound applications than DNA methylation data analysis.

Fortunately, many research experience from the gene expression analysis and the single-cell RNA-seq based or bulk RNA-based bioinformatics deconvolution methods can be useful for the research in DNA-methylation-based deconvolution, e.g., the DNA-methylation-based deconvolution method Emeth [6.NP21] is inspired by the gene expression-based deconvolution ICeD-T[7.JASA19]. We mainly turn to review and compare these two categories of different data-based cell type deconvolution methods in recent years later in this review later. Since these two categories of different data-based cell type deconvolution methods may will promote each other's development in the future.

Here we firstly present a total Tab.2 of the methods we are going to review later in this literature:

Categories of Methods	Methods	Description	Link
Gene expression-based deconvolution methods	SCADIE	A Comprehensive method can simultaneous estimating cell type-specific gene expression profiles and cell type proportions, even also can conduct the group cell type-specific differential expression testing	[3.GB22]
	CIBERSORTx	A support vector regression machine learning method which can provide detailed portraits of tissue composition without physical dissociation, antibodies or living material.	[8.NB19]
	csSAM	A early deconvolution method which estimates cell/tissue specific signatures from know proportions using SAM	[9.NM10]
	TOAST	A method provides a rigorous statistical framework with the pre-request that we know the mixing proportions, a variety of cell-type specific inferences can be drawn directly from testing different linear combinations of the linear model coefficients.	[10.Bio19]
	MuSic	A reference-based method which priorly utilizes cell-type specific gene expression from single-cell RNA sequencing (RNA-seq) data to characterize cell type compositions from bulk RNA-seq data in complex tissues.	[12.NC19]
	CMP	A reference-based method uses linear regression to estimate the expression abundance of reference cells in the given bulk samples. Priorly CPM constructs its reference collection from scRNA-seq profiles derived from one or a few relevant samples,	[11.NM19]

	DWLS	A method employs a weighted least squares method to estimate cell-type proportions	[13.NC19]
	ICeD-T	A method employs a mixture of regression model to identify those genes whose expression in a tissue sample is inconsistent to deal with aberrant genes.	[7.JASA19]
DNA methylation-based deconvolution methods	QP	A linear regression method with quadratic programming to impose the constraint that the regression coefficients are none-negative,	[19.BMC12]
	RLS(or Epidish),	A combined algorithm RLS(Epidish) which uses the new framework DHS data also they develop in a same research and robust partial correlations together for inference.	[20.BMC17],
	SVR(or MethylCIBERSORT)	A support vector regression machine learning method learned directly from CIBERSORT based on DNA methylation data	[21.NC18]
	EMeth	A reference-based method which partly learns from the gene-expression based method ICeD-T, aiming at overcoming the two disadvantages about the inaccuracy and unavailability of the reference cell type-specific DNA methylation database	[6.SR21]

Table 2. A Total Table of the Deconvolution Methods in this Review

Gene Expression-based Deconvolution Methods

Comparing with bulk RNA-seq and other gene expression-based data, in an individual sample given, although scRNA-seq usually has significant better performance in dissecting the heterogeneity of cellular compositions, as mentioned before, due to the high cost of scRNA-seq, bulk RNA-seq is still the main dataset used currently to further develop the cell type deconvolution methods.

In recent years, lots of deconvolution approaches are available for deconvoluting the compositions/proportions information of specific cell types from obtained bulk RNA-seq and other gene expression-based data.

Such as SCADIE[3.GB22], CIBERSORTx[8.NB19], csSAM[9.NM10], TOAST[10.Bio19], CMP[11.NM19], MuSiC[12.NC19], DWLS[13.NC19], ICeD-T[7.JASA19], many more earlier deconvolution methods can be viewed in the literature review [14.COI13] with specific applications in immune system, which is easy to see the necessity for us to apply deconvolution methods to immune system, since analyzed samples from immune system are often heterogeneous with respect to cell subsets which can mislead result interpretation. Now in this section, let's go through these methods' main theory principals:

Before the CIBERSORTx in [8.NB19], they previously developed an approach for digital cytometry, called CIBERSORT[15.NM15], as introduced in [6.NP21], this method enables estimation of cell type abundances from bulk tissue transcriptomes. The core of CIBERSORT is a support vector regression where the response variable is the gene expression from bulk tissues and each covariate corresponds to the gene expression from one cell type, which are usually estimated from external reference

samples. The good performance of CIBERSORT is in part due to the fact that the objective function of a support vector regression is robust to the noise in the data. When it comes to CIBERSORTx, as it's developed in[8.NB19], it's a machine learning method that extends CIBERSORT framework to infer cell-type-specific gene expression profiles without physical cell isolation. By this kind of extension, it brings new functionalities for cross-platform data normalization and in silico cell purification. Especially, the latter allows the transcriptomes of individual cell types to be extracted from bulk RNA admixtures without physical isolation. As a result, changes in cell-type-specific gene expression can be inferred without cell separation or prior knowledge. By leveraging cell type expression signatures from single-cell experiments or sorted cell subsets, CIBERSORTx can provide detailed portraits of tissue composition without physical dissociation, antibodies or living material.

CMP[11.NM19] provides an advantageous alternative to existing deconvolution approaches, particularly in providing a fine-resolution mapping. Different from some existing deconvolution methods, CMP uses linear regression to estimate the expression abundance of reference cells in the given bulk samples. While mainly based on the bulk samples, priorly CPM constructs its reference collection from scRNA-seq profiles derived from one or a few relevant samples, and then exploits this collection to infer cell composition within additional, bulk-profiled samples. Just like CMP, as developed in MuSiC[12.NC19], it is also a method that priorly utilizes cell-type specific gene expression from single-cell RNA sequencing (RNA-seq) data to characterize cell type compositions from bulk RNA-seq data in complex tissues. It weights the genes exhibiting cross-subject and cross-cell consistency to transfer cell-type-specific gene expression profile across different datasets.

Besides the (linear) regression method CMP[11.NM19], another recent (log) regression method for gene expression-based deconvolution is ICeD-T[7.JASA19], which models gene expression by a log-normal distribution. As introduced in [6.NP21], the advantage of log is that when we evaluate the loss function, the log-scale gene expression variance is much more stable than in linear scale. As another gene expression data-based method, DWLS[13.NC19] employs a weighted least squares method to estimate cell-type proportions. ICeD-T also employs a mixture of regression model to identify those genes whose expression in a tissue sample is inconsistent ,e.g., inconsistent cell type-specific gene expression between purified reference samples and tumor samples, with the deconvolution model and ICeD-T is able to automatically identify aberrant genes whose expression are inconsistent with the deconvolution model and down-weights their contributions to cell type abundance estimates. The same data inconsistent problem also occurs in DNA methylation-based deconvolution methods, based on the thoughts of ICeD-T, it will later inspire the DNA methylation-based deconvolution method EMeth[6.SR21] in the next section.

When it comes to TOAST[10.Bio19], based on a general and flexible linear model which covers many other existing methods, they provide a rigorous statistical framework with the pre-request that we know the mixing proportions by experimental measure or computationally estimated by a number of existing methods. Under their model parameterization, the method provides great flexibility for detecting csDE/csDM.

As they introduce, a variety of cell-type specific inferences can be drawn directly from testing different linear combinations of the linear model coefficients. Another method which relies on the known proportions is csSAM, it can be used to estimate cell/tissue specific signatures. The method csSAM[9.NM10], i.e., the short for cell type-specific significance analysis of microarrays, is one of the earliest method focused on analyzing differential gene expression for each cell type in a biological sample from microarray data and relative cell-type frequencies. In [10.Bio19], it shows that TOAST provides superior computational performance since it is directly based on linear regression while the method csSAM relies on some permutation procedure which leads to the simulation results that csSAM is much more computationally demanding than TOAST.

Recently, built on all possible existing deconvolution methods, which includes all previous mentioned methods like CIBERSORTx, csSAM, TOAST, CMP, MuSiC, DWLS, ICeD-T. SCADIE[3.GB22] is an very unique iterative algorithm that can be used to simultaneously estimate cell type-specific gene expression profiles and estimate the cell type proportions, as well as performs cell type-specific differential expression analysis, i.e., DEGs at the group level. SCADIE considers two groups of bulk RNA-seq samples, they aim to simultaneously estimate group-specific cell type-specific gene expression matrix Ws and cell type-specific proportion matrix Hs (s=1,2), where s is the index of the two groups of bulk samples in a two-group comparison setting, thus to accurately infer cell type-specific differentially expressed genes (DEGs) as well as cell type proportion changes. It takes bulk gene expression along with a common signature matrix or initial cell type proportions as input and then estimates group specific Ws and Hs. As introduced in [3.GB22], under the assume that the groups cell type-specific W1 , W2 are reasonably similar but not exactly the same while the theory cell type-specific W1 , W2 are strictly similar, it takes bulk gene expression along with a common signature matrix or initial cell type proportions as input and then estimates group specific Ws and Hs(s=1,2). Then it is possible to initialize with the same W in theory and use an iterative algorithm NNLS (non-negative least squares) to search for optimal group-specific Ws, as well as finally perform hypothesis tests by calculating their z-score based on the standard errors of their difference $\sum_{ij} W_1(i,j) - W_2(i,j)$ and then obtaining a p-value for testing differential expression to identify cell type-specific DEGs. Through comprehensive simulation and real data analyses, they demonstrate that SCADIE is simultaneously capable of identifying cell type-specific DEGs between Ws and maintaining high accuracy in estimating Hs.

DNA Methylation-based Deconvolution Methods

As mentioned in Gene-expression-based deconvolution methods, there are two classes of methods studying cell type deconvolution methods using gene expression data, reference-based methods and reference-free methods, the previous methods rely on the known of some extra information like the cell type proportions while the later one doesn't need. So does in DNA-methylation-based deconvolution methods.

Reference-free DNA-methylation-based deconvolution methods although do not require the reference of cell type-specific DNA methylation data, as an price of reducing

the demand of information, this class of methods' main goal is just to account for the variation of cell type composition in the association analysis of DNA methylation, and the parameters about the bulk samples estimated from these methods are often only linear combinations of cell type proportions instead of getting the cell type proportions parameters directly, such the methods in developed [16.Bio14], [17.GB18],[18.GB19], [27.GB19].

Corresponding to the regression model in gene-expression based deconvolution methods and remind that both the gene expression and the DNA methylation of cells vary across cell types, when it comes to reference-based methods, we can use DNA methylation of multiple CpGs in a tissue sample as the response variable to replace the total gene expression matrix W, while at the same time we use the DNA methylation of these CpGs in a cell type as each covariate to replace the cell type-specific gene expression matrix Wi (i=1,2…) to build the new model. This direct corresponding explains why many DNA methylation based deconvolution methods can learn from existing gene expression based deconvolution methods.

However, each coin has two sides, though reference-based can estimate more specific information like cell type specific proportions than reference-free methods, however, the accuracy and availability of the reference cell type-specific DNA methylation sometimes can't be certainly guaranteed. Specifically speaking, for some CpGs, the DNA methylation in the reference samples may not accurately obtained from the cell type-specific DNA methylation in a reference tissue sample. While in other cases, reference may not be available for all cell types, e.g., when considering tumor immune microenvironment, we often have the reference for immune cell types instead of tumor cells.

Now we go to mainly review some reference-based DNA-methylation-based deconvolution methods, which includes QP[19.BMC12], RLS(or Epidish) [20.BMC17], SVR(or MethylCIBERSORT)[21.NC18], EMeth[6.SR21]:

In the earliest method QP[19.BMC12], their proposed method resembles regression calibration, which is a linear regression method with quadratic programming to impose the constraint that the regression coefficients are none-negative, where they assume a methylation signature to be a high-dimensional multivariate surrogate for the distribution of human white blood cell populations, i.e. the immune profile.

As for RLS(or Epidish), actually in the whole paper[20.BMC17], as introduced in their research, they do four significant things:

(I) Firstly, as an reference-based method, DNA methylation database comes first, so they firstly propose a novel framework for reference-based inference for leveraging cell-type specific DNAse Hypersensitive Site (DHS) information from the NIH Epigenomics Roadmap to construct an improved reference DNA methylation database.

(II) Secondly, based the previous DHS new framework, by using this framework they compare a widely used state-of-the-art reference-based algorithm, i.e., constrained projection, with two non-constrained approaches including CIBERSORT (Of course, this is a gene-expression data based method, which indicates again that gene expression-based method may inspire DNA

methylation-based methods) and a method based on robust partial correlations.

(III) Thirdly, they conclude that the widely-used constrained methods, no matter DNA methylation based or gene expression-based methods, projection technique may not always remain optimal which implies that more reference information doesn't mean more good functions. Actually, they find that the reference-free method based on robust partial correlations is generally more robust across a range of different tissue types and for realistic noise levels.

(IV) Finally, inspired by the robust liner regression method and the CIBERSORT, they develop a combined algorithm RLS(Epidish) which uses the new framework DHS data and robust partial correlations together for inference, the method RLS(Epidish) in fact replaces linear regression in QP with robust linear regression (R function MASS/ rlm) which uses a weighted loss function so that the data points with larger residuals have smaller weights.

Also motivated by the success of gene expression-based deconvolution method CIBERSORT in gene expression decomposition, in [21.NC18], they also employed SVR (Support Vector Regression) to estimate cell type composition using DNA methylation data, the main difference is that they use DNA methylation data to replace gene expression data. Naturally, this method is reasonable to be named as MethylCIBERSORT.

EMeth[6.NP21] is a reference-based method which aims at overcoming the two disadvantages mentioned before about the inaccuracy and unavailability of the reference cell type-specific DNA methylation database, these two limitations of DNA methylation database also partly explain why some reference-based methods with more extra information actually perform less robust than some reference-free method like RLS. As introduced in their research, they firstly use an EM (Expectation-Maximization) DNA methylation data based algorithm for parameter estimation.

Then they use following two tracks to overcome the two limitations:

Firstly, when facing the limitation of inaccurate reference, motivated again by one of the gene expression-based methods ICeD-T in [7.JASA19], they adopt a similar mixture of regression approach: for some CpGs, instead of considering all the models of each CpG with a single distribution for regular/consistent CpGs. EMeth actually models the observed DNA methylation of each CpG by a mixture distribution with one component for regular/consistent CpGs and the other component for aberrant CpGs, the aberrant CpGs means that their DNA methylation are inconsistent with what is expected from the deconvolution model, a note here is that the standard EM algorithm indeed can estimate the parameters of this mixture of regression model. By optimizing and reducing the possible inconsistent error, EMeth can automatically down-weigh the contributions of the aberrant CpGs on cell type deconvolution to approximate and remain the regular/consistent CpGs.

Secondly, when facing the problem of unavailability of the reference, actually the framework of EMeth take this unfortunate situation into consider and additionally includes a special cell type without methylation reference, but with known cell type proportions. In fact, EMeth is able to estimate the DNA methylation of this special cell

type. This special cell type is often the case for tumor tissues where tumor purity is known but DNA methylation in tumor cells is unknown, in fact, the mentioned technical track to include this unavailable additional special cell type term is motivated by cancer studies where tumor purity can be estimated from DNA copy number data, or even methylation data itself [23.GB17] .

The unknown methylation level of this special unavailable cell type can in fact be estimated by borrowing information across tissue samples. Specifically speaking, for each CpG, just like the previous model that each group's observed gene expression is the product of the group proportion and its group gene expression in [3.GB22], the expected contribution of this special cell type to the observed methylation is proportional to the product of its proportion in a tissue sample and its methylation. Therefore, as suggested in Tab.1, finally its methylation can be estimated by a regression using the methylation data of this CpG across tissue samples, where the response variable is the observed methylation, and the covariate is the proportion of this special cell type.

Simulations Results and Comparisons Review

Among these recent methods introduced in this review both for DNA-methylation-based data and gene-expression-based data, EMeth and SCADIE are two methods that have made many substantial and verified simulations and real data results, as well as some direct comparisons with some other methods. In this section, we will cite and review the simulations and comparisons results from[3.GB22] and [6. NP21] to go through the functions and get a sense of the performance strength of these methods within given some bulk samples or simulated data:

Firstly, we go to review part of the simulations and comparisons results in [3.GB22], all the results are cited from its original research:

In their original research, aiming at evaluating of different methods when using in silico mixtures of cell type-specific DNA methylation data, for each individual, they simulated a mixture by linearly combining cell type-specific data, followed by adding Gaussian noise in M-value scale for methylation data and in log-scale for expression data. The only parameter in this simulation study is the variance of the Gaussian noise.

Although EMeth is a DNA methylation-based method, they also compare it whit one classical gene expression-based method, the CBERSORT[15.NM15]/ CBERSORTx. To achieve this different data-based methods comparison goal, they use the same mixture proportions on both methylation and gene expression data so they are able to compare the cell type proportion estimates from these two types of data. In detail, they used the data from 56 individuals to construct the reference data to get the simulated prior knowledge of cell type-specific DNA methylation. Then they use the remaining 68 individuals to generate mixture bulk samples to simulate the estimation process of the bulk samples' cell type-specific proportions. To simulate the real situation, they also additionally added the same level of noise to both expression and methylation data. Their plots and graphs were all generated using R version 3.6.2[24.RCT20].

The simulation results show that, Both EMeth and RLS have accurate estimation results, which has a correlation with true cell type proportions as high as 0.95 and RMSE around 10^{-3} for each cell type while EMeth and RLS consistently outperform comparing to other methods like LS, QP, SVR and CIBERSORT(Fig. 1), where the method LS is the simplest ordinary linear regression which minimizes residual sum squares of the model fit, i.e., least squares or LS. In Fig.1, the researchers conduct some evaluation of the previous mentioned 6 different methods when using in silico mixtures. The upper row bar graphs show the correlation between the estimated proportions and the true proportions for each cell type and each method, in this simulation, they test 3 cell types which are Tcell, Monocyte and Neutrophil. Clearly, when the correlation signed on y-axis of the corresponded method is more closed to 1, the estimation is more accuracy.

While the lower row bar graphs display the rooted MSE, when the rooted MSE signed on y-axis is more closed to 0, the corresponded method is more accuracy. As for the same level part of noise, in Fig.1's experiments, they set the noise level parameter c as 5 to better demonstrate the difference across methods in a more real situation, and the conclusions are the same for other values of c.

Additionally, notice that this simulation also indicates that DNA methylation-data based method may perform better than gene expression-based data, though, EMeth is a reference-based method while the CIBERSORT is a reference-free method, as introduced in their research, the estimates by EMeth and RLS based on DNA methylation data are more accurate than the estimates by CIBERSORT based on gene expression data.

What's more, observing in another angle, the significantly better performances of EMeth and RLS are more highlight by their exam on the cell type proportion estimates for each specific cell type across all the individuals (Fig. 2). In Fig.2, they consider the graphs of CD4 T cells, the y-axis is the true proportion of CD4 T cells for all 68 samples and the x-axis is the proportion estimates by six different methods. When the data point for a given individual is more closed to the line y=x, then the estimation for this individual is more accurate. For the case of CD4 T cells, we can also observe from Fig.2 that EMeth and RLS give very accurate estimates from methylation data and other methods provide reasonable but less accurate estimates

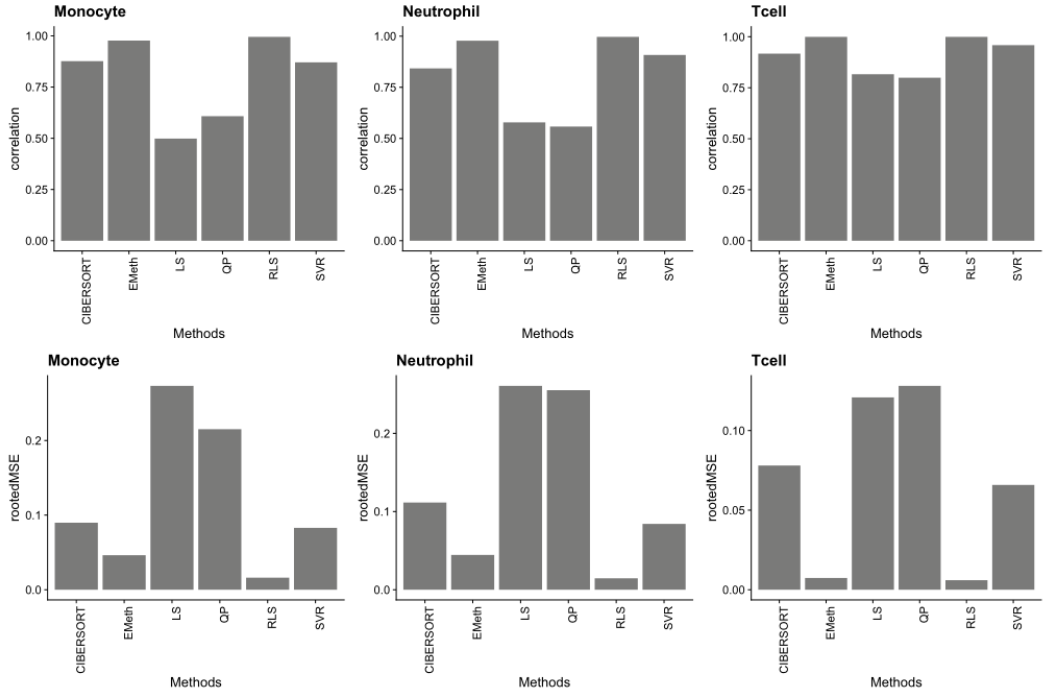


Figure 1: Evaluation of different methods using in silico mixtures. The upper row shows the correlation between the estimated proportions and the true proportions for each cell type and each method. The lower row displays the RMSE. The noise level parameter c were set as 5 to better demonstrate the difference across methods, and the conclusions are the same for other values of c.

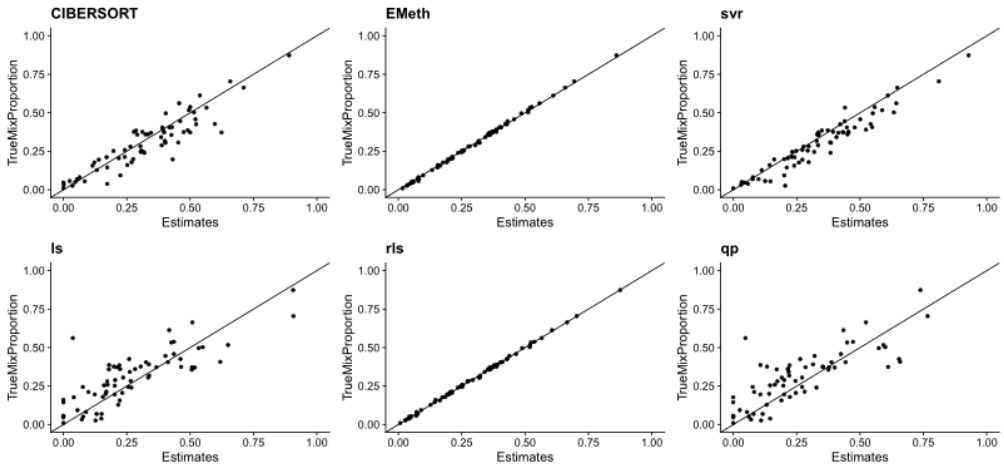


Figure 2: Evaluation of cell type proportion estimates for CD4 T cells using in silico mixtures. The y-axis is the true proportion of CD4 T cells for all 68 samples and the x-axis is the proportion estimates by six different methods. EMeth and RLS give very accurate estimates from methylation data and other methods provide reasonable but less accurate estimates

As for gene expression-based method SCADIE, it's a powerful, comprehensive method with lots of functions in cell type specific analysis, which includes cell-type-specific proportion estimation, cell type-specific gene expression profiles estimation

and a powerful function on cell type-specific differentially expressed genes (DEGs) testing between groups. It can also generally help to improve the estimates from other methods. Here we take the cell-type-specific proportion estimation as an example to review the performance of these gene-expression based deconvolution methods:

In [3.GB22], to evaluate and compare the proposed SCADIE method performance on cell type proportion estimates with other methods, they benchmarked SCADIE against four deconvolution algorithms, which includes CIBERSORTx[8.NB19], MuSiC[12.NC19], DWLS[13.NC19], and a naive version of SCADIE by directly using NNLS(Non-negative Least Squares) in updating W.

They tested these four methods on the following two datasets:A simulated data set, a pseudo-bulk data set[25.Na17], and a bulk microarray data with known cell type proportions[26. NM10], which includes three sub kinds of datasets, simulation datasets, mouse ISC pseudo bulk, mouse bulk datasets.

As two quantitative ways to compare these methods, they both used two metrics to evaluate the accuracy of the estimated Hs(s=1,2): K-L divergence and root-mean-squared error (RMSE). A Method performs better when these two metrics indexes are lower than another method.

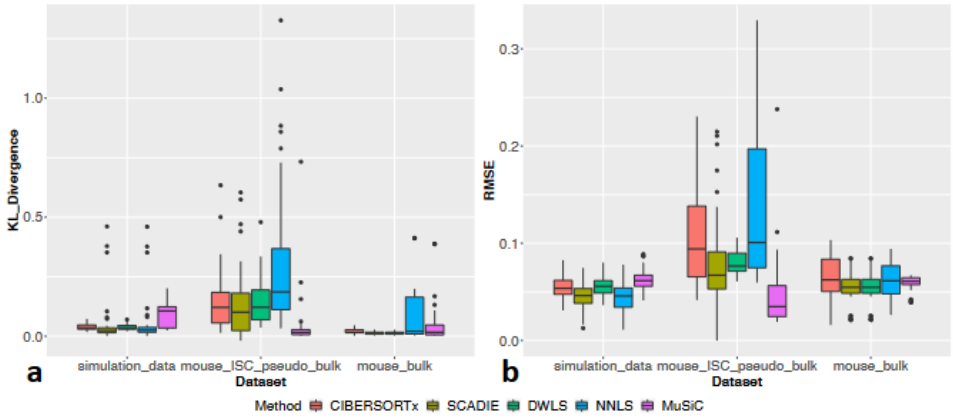


Figure 3[3]: Benchmark cell type proportion estimations from SCADIE against DWLS, CIBERSORTx, MuSiC, and the naive iterative procedure with NNLS W-update: a. K-L Divergence between H and the ground truth proportions across three data sets, SCADIE and NNLS iteration's Hs were from the final iteration output, and Hs of DWLS and CIBERSORTx were directly from deconvolution; b. Same results as a) but measured by root-mean-square error (RMSE), the result patterns are consistent with those in K-L Divergence;

As we can observe, SCADIE showed equal or better accuracies than the other four methods only except in the mouse ISC pseudo bulk dataset. In fact, ISC pseudo bulk dataset is a single cell data while MuSiC especially suits for , since it's a cell type deconvolution method which utilizes cross-subject scRNA-seq to estimate cell type

[3] **This boxplot is cited from Additional File 1: Supplementary Fig. S3ab in [3.GB22]**

proportions in bulk RNA-seq data. So from the simulated performance of SCADIE in estimating cell type proportion, it has a more sustainable and better performance in most datasets than other gene expression based deconvolution methods.

Conclusions

In recent years, since both the gene expression and the DNA methylation of cells vary across cell types, more and more methods based on these two kinds of Datasets focused on cell type deconvolution with varied functions, efficiency and theory principles have been developed.

In functions, some methods are more focused on cell type-specific compositions/proportions estimation, some are more focused on cell type-specific gene expression while other methods may be more focused on s cell type-specific differential expression analysis. Methods like SCADIE may also have very comprehensive functions on cell type-specific analysis comparing to some other single function methods.

In efficiency, different methods may have difference efficiency and performance when they are dealing with different type specific analysis purpose, like MuSiC especially suits for single cell count data when it compares to other methods. ICeD-T suits for some gene expression data with aberrant patterns like data from bulk tumor samples. However, totally speaking some methods are significantly more powerful and more comprehensive than other methods, e.g. SCADIE generally performs well than DWLS, CIBERSORTx, MuSiC in cell type specific-proportions estimation and other functions. EMeth and RLS perform better than other DNA methylation-based deconvolution methods and a gene-expression based method CIBERSORTx, which indicates that DNA methylation-based method has the potential to perform well than gene expression-based. In this review, SCADIE performs better than other gene expression-based methods except in single cell count data when it compares to MuSiC while EMeth and MethylCIBERSORT(SVR) [21.NC18] perform better than other DNA methylation-based methods.

In theory principles, different methods have fruitful and varied theory principles, for example, some methods are reference-based method, like MuSiC will firstly perform scRNA-seq on a few samples which is an efficient and cost-effective way to generate the cell-type-specific gene expression profile as the reference. However, in contrast methods like DeCompress[28.NAR21] is a reference-free or semi-reference-free gene expression-based method. So does the side of DNA methylation-based methods, like EMeth is a reference-based method while there are also many DNA methylation-based and also reference-free based methods like the methods developed in [16.Bio14], [17.GB18], [18.GB19], [27.GB19]. Also, among these methods, different set-up regression models and different applied corresponding optimal techniques are very rich and varied, e.g., which includes but not only includes Ordinary Line Regression, Log-Line regression, Weighted Least Squares Method, Permutation Procedure, Support Vector Regression, various Machine Learning Methods, Non-negative Least Squares Optimal, Robust Linear Regression, Linear Regression with some Constraints,

Quadratic Programming…

DNA methylation-based Deconvolution Methods and Gene expression-based Deconvolution Methods all have its benefits and disadvantages, generally speaking, DNA methylation-based methods' data source is usually more stable than gene expression and DNA methylation is easier to measure in FFPE tissues or formalin-fixed paraffin-embedded[4.BM10], while some gene-expression data like scRNA-seq data usually has high cost and complexity. On the other hand, gene expression-based deconvolution methods are more fruitful and currently have more available research than DNA methylation-based deconvolution methods. So DNA methylation-based methods in some cases learn from the gene expression-based method, like the EMeth learns from ICeD-T in [7.JASA19] and the MethylCIBERSORT [21.NC18] learns from CIBERSORT[15.NM15]. All in all, both of these two kinds of different data based methods are powerful tools to realize the purpose of cell type-specific deconvolution methods and they have been rapidly developing in recent years with more coming new methods in the future.

References

1. Jaffe, A.E., Irizarry, R.A.: Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol.15(2), 1–9 (2014).
2. Sinha BK, Tokar EJ, Bushel PR. Elucidation of Mechanisms of Topotecan-Induced Cell Death in Human Breast MCF-7 Cancer Cells by Gene Expression Analysis. Front Genet. 2020 Jul 17; 11: 775.
3. Tang, D., Park, S., & Zhao, H. (2022). SCADIE: Simultaneous estimation of cell type proportions and cell type-specific gene expressions using SCAD-based iterative estimating procedure. Genome Biology, 23, 1-23.
4. Thirlwell, C., Eymard, M., Feber, A., Teschendorff, A., Pearce, K., Lechner, M., Widschwendter, M., Beck, S.: Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the illumine infinium humanmethylation27 beadchip. Methods 52(3), 248–254 (2010)
5. Yunjin Li, Qiyue Xu, Duo jiao, Wu, Geng Chen, Front. Cell Dev. Biol., Exploring Additional Valuable Information From Single-Cell RNA-Seq Data, 01 December 2020 Sec. Molecular and Cellular Pathology https://doi.org/10.3389/fcell.2020.593007
6. Hanyu, Z., Ruoyi, C., Dai, J., & Sun, W. (2021). EMeth: An EM algorithm for cell type decomposition based on DNA methylation data. *Scientific Reports (Nature Publisher Group), 11*(1) doi:https://doi.org/10.1038/s41598-021-84864-9
7. Wilson, D. R., Jin, C., Ibrahim, J. G. & Sun, W. Iced-t provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. J. Am. Stat. Assoc. https ://doi.org/10.1080/01621 459.2019.16548 74 (2019).
8. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotech nol. 2019;37(7):773–82
9. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM,

Davis MM, Butte AJ. Cell type–specific gene expression differences in complex tissues. Nat Methods. 2010;7(4):287–89

10. Li Z, Wu Z, Jin P, Wu H. Dissecting differential signals in high-throughput data from complex tissues. Bioinformatics.2019;35(20):3898–905.
11. Frishberg, A., Peshes-Yaloz, N., Cohn, O., Rosentul, D., Steuerman, Y., Valadarsky, L., et al. (2019). Cell composition analysis of bulk genomics using single-celldata. Nat. Methods 16, 327–332. doi: 10.1038/s41592-019-0355-5
12. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10(1):1–9
13. Tsoucas, D., Dong, R., Chen, H. D., Zhu, Q., Guo, G. J., and Yuan, G. C. (2019). Accurate estimation of cell-type composition from gene expression data. Nat. Commun. 10:2975.
14. Shai S Shen-Orr, Renaud Gaujoux Computational deconvolution: extracting cell type-specific information from heterogeneous samples Current Opinion in Immunology Volume 25, Issue 5, October 2013, Pages 571-578
15. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 12, 453–457 (2015)
16. Houseman, E.A., Molitor, J., Marsit, C.J.: Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics 30(10), 1431–1439 (2014).
17. Rahmani, E., Schweiger, R., Shenhav, L., Wingert, T., Hofer, I., Gabel, E., Eskin, E., Halperin, E.: Bayescce: a bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. Genome Biol. 19(1), 141(2018).
18. 11. Li, Z., Wu, H.: Toast: improving reference-free cell composition estimation by cross-cell type differential analysis. Genome Biol.20(1), 1–17 (2019).
19. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., Kelsey, K.T.: DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform 13(1), 86(2012). https ://doi.org/10.1186/1471-2105-13-86.
20. Teschendorff, A.E., Breeze, C.E., Zheng, S.C., Beck, S.: A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. BMC Bioinform 18(1), 105–105 (2017). https ://doi.org/10.1186/s1285 9-017-1511-5.
21. Chakravarthy, A., Furness, A., Joshi, K., Ghorani, E., Ford, K., Ward, M.J., King, E.V., Lechner, M., Marafioti, T., Quezada, S.A.,et al.: Pan-cancer deconvolution of tumour composition using DNA methylation. Nat. Commun. 9(1), 3220 (2018).
22. Newman, A., Liu, C., Green, M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 12, 453–457 (2015). https://doi.org/10.1038/nmeth.3337
23. Zheng, X., Zhang, N., Wu, H.-J., Wu, H.: Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. Genome Biol. 18(1), 1–14 (2017).
24. R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
25. YanKS, Janda CY, Chang J, Zheng GX, Larkin KA, Luca VC, Chia LA, MahAT, Han A, Terry JM, etal. Non-equivalence of wnt and r-spondin ligands during lgr5+ intestinal stem-cell self-renewal. Nature. 2017;545(7653):238–42.
26. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ. Cell type–specific gene expression differences in complex tissues. Nat

Methods. 2010;7(4):287–89.

27. Li Z, Wu H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. Genome Biol. 2019 Sep 4;20(1):190. doi: 10.1186/s13059-019-1778-0. PMID: 31484546; PMCID: PMC6727351.
28. Bhattacharya A, Hamilton AM, Troester MA, Love MI. DeCompress: tissue compartment deconvolution of targeted mRNA expression panels using compressed sensing. Nucleic Acids Res. 2021 May 7;49(8):e48. doi: 10.1093/nar/gkab031. PMID: 33524140; PMCID: PMC8096278.

Appendix A: More Beyond Cell Type Deconvolution Methods: Some Other Cell Type-Specific Analysis Applications

Cell type specific analysis in recent year has become a highly and rapidly developed subject with a lot of sub-branches, technical interactions and various interesting applied applications in Bioinformatics, Cell Biology, Biostatistics, Immunology, Cancer Research, Clinical, Genetics…

Beyond the deconvolution methods in cell type-specific analysis, there are many other interesting applications and aspects in cell type specific analysis, here we will take a brief look at some other specific applications and methods in the big family of cell type-specifc analysis, for example, identify cell-type–specific APA genes in scRNA-Seq data, Classifying Cell Type-specific Enhancers, Transcript Unit-calling Algorithms and Network Analysis. The references in this appendix are sort out in the order of years and included in the Appendix B, a time table of partial methods and technology in cell type-specifc analysis

Identify cell-type–specific APA genes in scRNA-Seq data

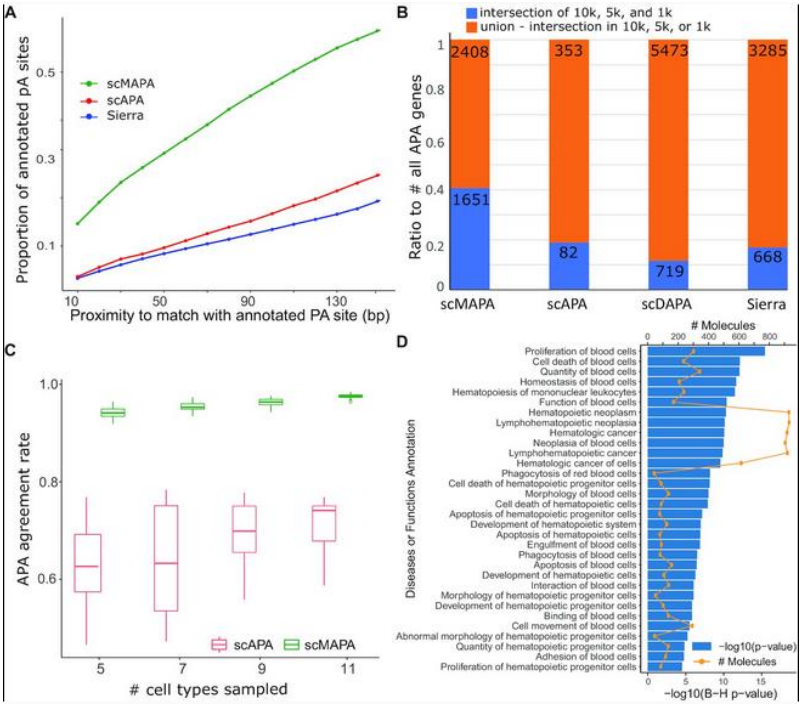
Single-cell RNA-seq (scRNA-seq) is a powerful tool for studying gene expression in single cells. Among its application, alternative polyadenylation (APA) plays a key post-transcriptional regulatory role in mRNA stability and functions in eukaryotes while single cell RNA-seq (scRNA-seq) is a powerful tool to discover cellular heterogeneity at gene expression level. To identify cell-type–specific APA genes in scRNA-Seq data, several bioinformatic methods have been developed, such as scMPAP[Giga22], scDAPA[Bio20], Sierra[GB20] and scAPA[NAR19]:

scMPAP

scMPAP developed a combination of a computational change-point algorithm and a statistical model, single-cell Multi-group identification of APA (scMAPA).

To avoid the assumptions on the read coverage shape, scMAPA formulates a change-point problem after transforming the 3′ biased scRNA-Seq data to represent the full-length 3′-UTR signal. To identify cell-type–specific APA genes while adjusting for undesired source of variation, scMAPA models APA isoforms in consideration of the cell types and the undesired source.

In their novel simulation data and data from human peripheral blood mononuclear cells, scMAPA outperforms existing methods in sensitivity, robustness, and stability(Pic1.1). So scMAPA elucidates the cell-type–specific function of APA events and sheds novel insights into the functional roles of APA events in complex tissues.



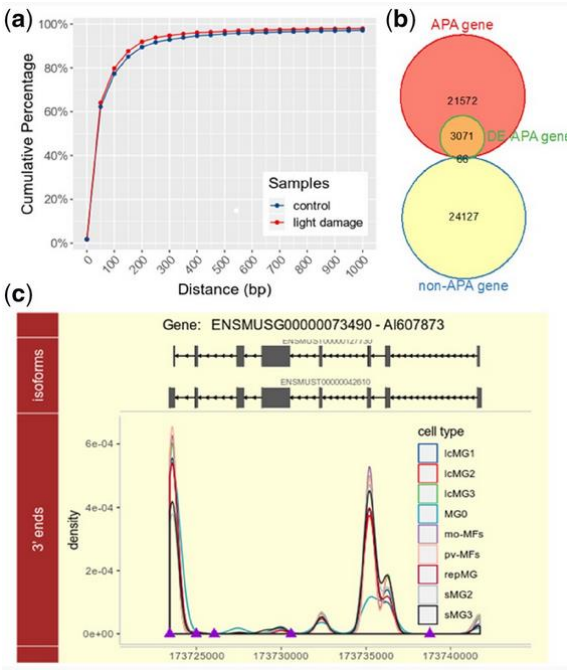
Pic1.1 Compare scMAPA with other methods

scDAPA

They developed a package scDAPA to detect and visualize dynamic APA from scRNA-seq data. We demonstrated its utilities through application to a real dataset by investigating the application of scDAPA on a scRNA-seq dataset of live microglia/macrophages from pooled neuroretinas of normal and light damaged mice generated by the 10× Genomics platform.(pic1.2)

scDAPA is a useful tool in studying APA at single cell resolution, and will broadly

extend the application scope of scRNA-seq data.



Pic1.2: An application example of scDAPA on a scRNA-seq dataset of neuroretinas from mouse.

Sierra

They present a computational pipeline, Sierra, that readily detects differential transcript usage from data generated by commonly used polyA-captured scRNA-seq technology, as an application, they validate Sierra by comparing cardiac scRNA-seq cell types to bulk RNA-seq of matched populations, finding significant overlap in differential transcripts. Sierra detects differential transcript usage across human peripheral blood mononuclear cells and the Tabula Muris, and 3 ′UTR shortening in cardiac fibroblasts.

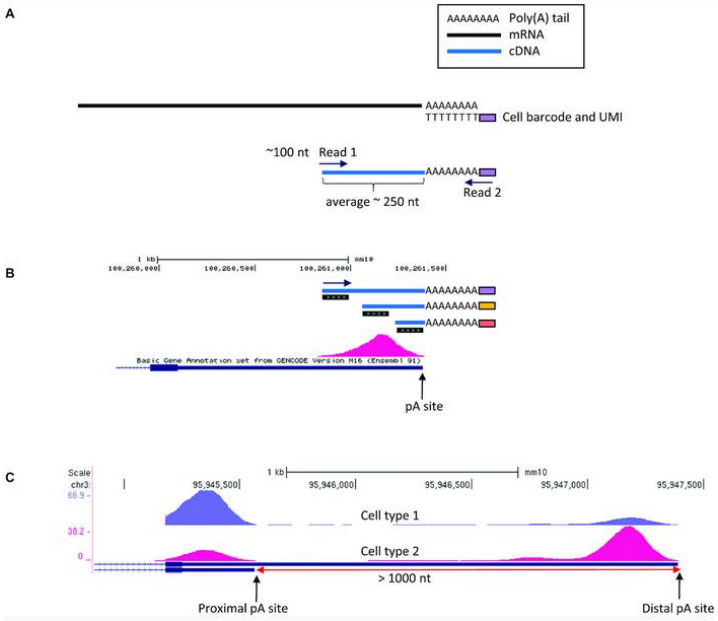
scAPA

Back to scRNA-seq data again, novel single-cell RNA sequencing (scRNA-seq) techniques allow molecular characterization of different cell types to an unprecedented degree.Especially, the most popular scRNA-seq protocols specifically sequence the 3′ end of transcripts.

Building on this property, they implemented a method for analysing patterns of APA regulation from such data. Analyzing multiple datasets from diverse tissues.

As an application, they identified widespread modulation of APA in different cell

types resulting in global 3′ UTR shortening/lengthening and enhanced cleavage at intronic pA sites(Pic1.3). Their results provide a proof-of-concept demonstration that the huge volume of scRNA-seq data that accumulates in the public domain offers a unique resource for the exploration of APA based on a very broad collection of cell types and biological conditions.



Pic1.3 Utilizing 3′ tag scRNA-seq data for the study of APA.

Classifying Cell Type-specific Enhancers

Cell type-specific enhancers, *cis*-regulatory elements that up-regulate gene transcription in a cell type, play a key role in determining the regulatory landscape of the human genome.

Predicting enhancers based on transcription factor binding sites (TFBS) was proposed because TFBS tend to be conserved over vertebrate evolution. However, there is uncertainty regarding the identification of TFBS from DNA sequences.

To ameliorate this challenge, direct sequence features such as *k*-mers (i.e., nucleotide sequences with a specified length) were then introduced to model enhancer prediction [GR11] [PLOS14]. However, these early studies did not achieve high prediction accuracy nor were they able to distinguish enhancers of different cell types.

SeqEnhDL

Fortunately, in recent years, deep learning technologies have gained greater popularity compared to conventional machine learning methods, and have been adapted in biomedical research to address complex research questions. Thus, deep learning can be more powerful in classifying enhancers.

So they propose SeqEnhDL, a deep learning framework for the classification of cell type-specific enhancers based on sequence features. To include interdependency and sequence information in the features of a DNA sequence, SeqEnhDL uses positional *k*-mer fold changes across each nucleotide position as its features. The effectiveness and advantages of SeqEnhDL are demonstrated based on the chromatin state segmentation data of nine cell types from the ENCODE project.(Pic2.1)

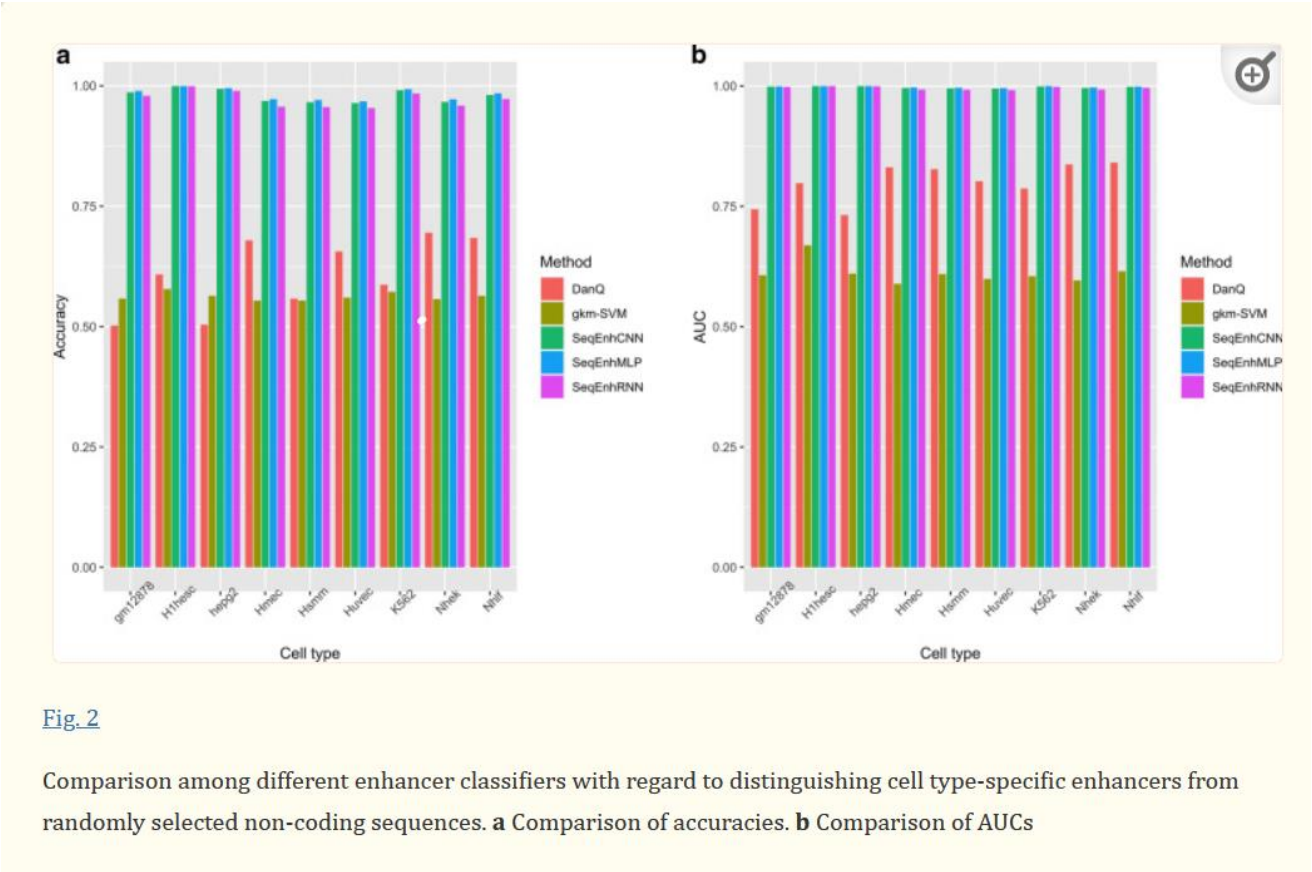


Fig. 2

Comparison among different enhancer classifiers with regard to distinguishing cell type-specific enhancers from randomly selected non-coding sequences. **a** Comparison of accuracies. **b** Comparison of AUCs

Pic 2.1 Compare Different Enhancer Classifiers

gkm-SVM

Oligomers of length k, or k-mers, are convenient and widely used features for modeling the properties and functions of DNA and protein sequences.

However, unfortunately, k-mers suffer from the inherent limitation that if the parameter k is increased to resolve longer features, the probability of observing any specific k-mer becomes very small, and k-mer counts approach a binary variable, with most k-mers absent and a few present once. Thus, any statistical learning approach using k-mers as features becomes susceptible to noisy training set k-mer frequencies once k becomes large.

So to address this problem, they introduce alternative feature sets using gapped k-mers, a new classifier, gkm-SVM, and a general method for robust estimation of k-

mer frequencies. To make the method applicable to large-scale genome wide applications, we develop an efficient tree data structure for computing the kernel matrix.

They show that compared to their original kmer-SVM and alternative approaches, our gkm-SVM predicts functional genomic regulatory elements and tissue specific enhancers with significantly improved accuracy, increasing the precision by up to a factor of two. They then show that gkm-SVM consistently outperforms kmer-SVM on human ENCODE ChIP-seq datasets, and further demonstrate the general utility of our method using a Naïve-Bayes classifier. Although developed for regulatory sequence analysis, these methods can be applied to any sequence classification problem.

SVM

Accurately predicting regulatory sequences and enhancers in entire genomes is an important but difficult problem, especially in large vertebrate genomes. With the advent of ChIP-seq technology, experimental detection of genome-wide EP300/CREBBP bound regions provides a powerful platform to develop predictive tools for regulatory sequences and to study their sequence properties.

They develop a support vector machine (SVM) framework which can accurately identify EP300-bound enhancers using only genomic sequence and an unbiased set of general sequence features. Moreover, they find that the predictive sequence features identified by the SVM classifier reveal biologically relevant sequence elements enriched in the enhancers, but they also identify other features that are significantly depleted in enhancers.

Transcript Unit-calling Algorithms

Global run-on coupled with deep sequencing (GRO-seq) provides extensive information on the location and function of coding and non-coding transcripts, including primary microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and enhancer RNAs (eRNAs), as well as yet undiscovered classes of transcripts. There are a few computational tools tailored toward this new type of sequencing data which are available.

groHMM

GroHMM[Bio15] is a complete pipeline for the accurate identification of the boundaries of transcriptional activity across the genome using GRO-seq data and the classification of these transcription units using a database of available annotations, which is provided as an R package in Bioconductor.

In addition, they describe novel metrics for the accuracy of transcription unit annotation, which show that groHMM substantially outperforms alternative approaches for identifying both coding and non-coding transcription units. To demonstrate the utility of their approach, they use groHMM to annotate four GRO-seq data sets derived

from cells representing a variety of different human tissue types, as well as non-mammalian cells. Their analyses using groHMM, a complete and useful tool for evaluating functional elements in cells, reveal new insights into cell type-specific transcription.

A systematic comparison of the performance between groHMM and two existing peak-calling methods tuned to identify broad regions (SICER and HOMER) favorably supports their approach groHMM on existing GRO-seq data from MCF-7 breast cancer cells.(Pic3.1&Pic3.2)

SICER

Based on the biological observation that histone modifications tend to cluster to form domains, SICER[Bio09] presents a method that identifies spatial clusters of signals unlikely to appear by chance. This method pools together enrichment information from neighboring nucleosomes to increase sensitivity and specificity. By using genomic-scale analysis, as well as the examination of loci with validated epigenetic states, they demonstrate that this method outperforms existing methods in the identification of ChIP-enriched signals for histone modification profiles. They demonstrate the application of this unbiased method in important issues in ChIP-Seq data analysis, such as data normalization for quantitative comparison of levels of epigenetic modifications across cell types and growth conditions.

HOMER

HOMER [MC10] is a method that identifies a sudden increase in GRO-seq signal to denote the start of a transcription unit. The signals are considered artificial spikes if they fail to last over a large distance.

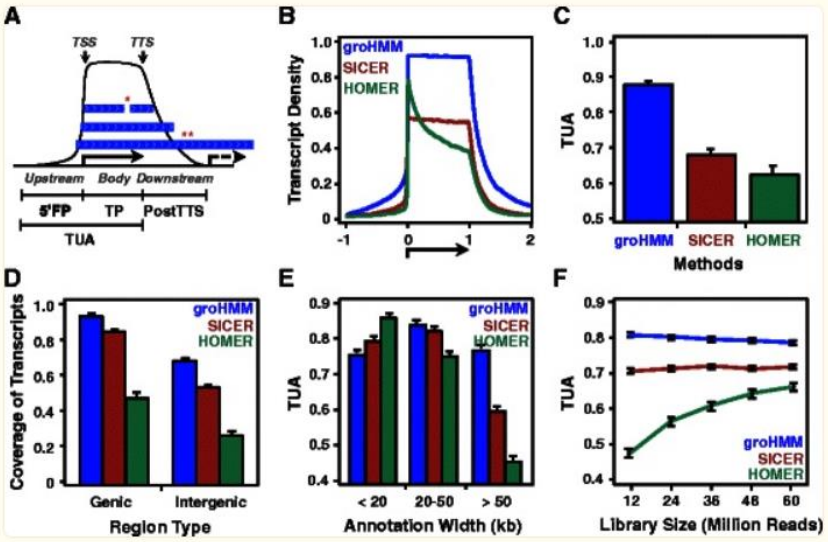
RSEG

They present the RSEG method [Bio11] for identifying epigenomic domains from ChIP-Seq data for histone modifications, which is an HMM-based tool for calling broad peaks of histone modifications from ChIP-seq data. In contrast with other methods emphasizing the locations of ‘peaks’ in read density profiles, our method identifies the boundaries of domains. RSEG is also able to incorporate a control sample and find genomic regions with differential histone modifications between two samples.

Vespucci

They present a novel algorithm [NAR13] for de novo transcript identification from GRO-sequencing data, along with a system that determines transcript regions, stores them in a relational database and associates them with known reference annotations.

As an application, they use this method to analyze GRO-sequencing data from primary mouse macrophages and derive novel quantitative insights into the extent and characteristics of non-coding transcription in mammalian cells.



(Pic 3.1 compare the performance of each method with groHMM)

Table 2

Optimal parameter values and error rates for each transcript-calling algorithm tested using GRO-seq data from MCF-7 cells

Method	Parameters	Optimal value	Number of transcripts	Median transcript length (bp)	Error		
					Merged annotation	Dissociated annotation	Rate
groHMM	-LtProbB (*T*)	350	29,639	7,750	1,956	745	0.065
	UTS (σ^2)	30					
SICER	windowSize	1,200	26,066	13,200	1,602	2,099	0.097
	gapSize	3,600 (3x)					
HOMER	minBodySize	2,500	25,542	4,240	731	1029	0.047
	bodyFold	12					

Open in a separate window

(Pic 3.2 compare the performance of each method with groHMM)

Network Analysis

Gene networks are rapidly growing in size and number, raising the question of which networks

are most appropriate for particular applications. Besides the recent SCINET [CS19] on the reconstruction of cell-type-specific interactomes, in [CS18], they evaluate 21 human genome-wide interaction networks for their ability to recover 446 disease gene sets identified through literature curation, gene expression profiling, or genome-wide association studies.

SCINET

They introduced SCINET [CS19], a computational framework that enables the reconstruction of cell-type-specific interactomes by leveraging single-cell transcriptomic data. By inferring and quantifying cell-type-specific gene interaction strengths, SCINET provides a cellular context to interpret molecular pathways and functional modules. SCINET can be used to contextualize disease-associated genes and their quantifiable influence in different cell types or conditions, to study potential mediators of functional interactions between cell types, or to assess the dynamics of interaction usage across developmental or pathological conditions.

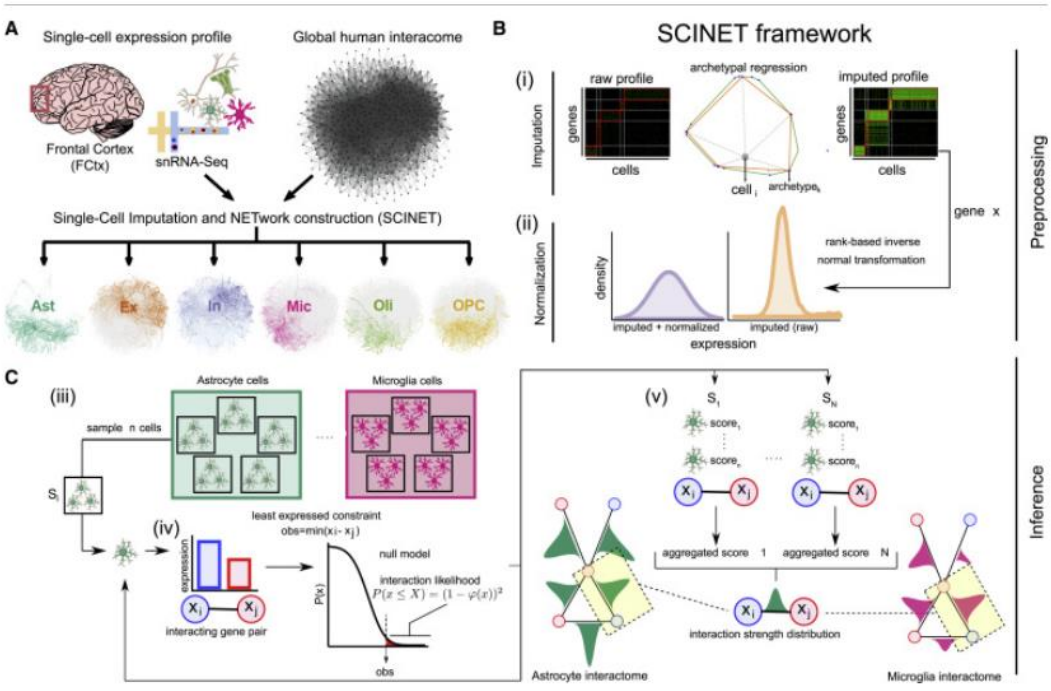
The core SCINET framework(Pic 4.1) is based on the following methodological developments:

(1) a decomposition method to interpolate values for missing observations in the scRNA-seq profile,

(2) a parametric approach to project heterogeneous gene expression distributions into a compatible subspace ,

(3) a statistical framework to measure the likelihood of gene interactions within each cell, and

(4) a subsampling approach to aggregate interaction likelihoods of individual cells, reduce noise, and estimate the underlying distribution and variability of interaction strengths within each cell-type population .



Download : Download high-res image (1MB) Download : Download full-size image

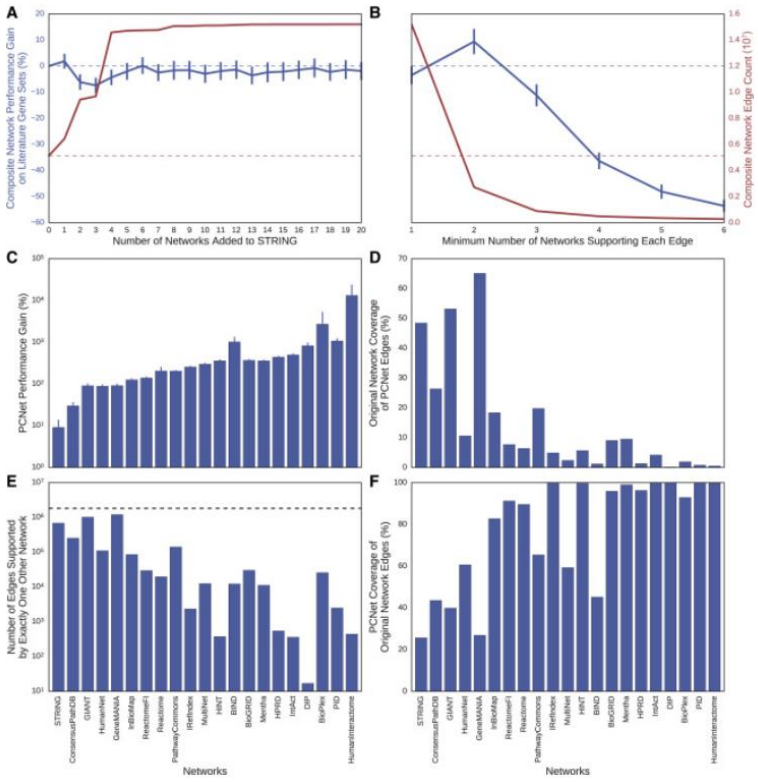
(Pic 4.1 SCINET framework)

PCNet

In [CS18], they evaluate 21 human genome-wide interaction networks for their ability to recover 446 disease gene sets identified through literature curation, gene expression profiling, or genome-wide association studies and also

The result shows that larger networks outperform smaller ones, as a general trend, supports the continued investment in high-throughput discovery of biological interaction networks. Now that, given the good performance of molecular networks that are large and inclusive, they next considered that these separate resources might be further improved by combining them to form a single composite network. They created a series of composite networks of decreasing size, by requiring interactions to be present in ever greater numbers of individual networks. By requiring a minimum of two networks supporting each interaction, the performance was significantly improved over the best individual network despite having a much smaller network size. This configuration was optimal, since further increasing the minimum number of supporting networks beyond two resulted in a degradation of performance.They call this optimal configuration the “Parsimonious Composite Network” (PCNet).

At the same time, they were able to derive a much smaller PCNet that outperformed a network twice its size on the literature gene set recovery tasks (Pic4.2). This observation suggests at least one straightforward method of contracting the size of a reference network without sacrificing performance: requiring multiple database support for interactions.



(Pic 4.2. Composite Networks Can Gain Performance Despite Smaller Size)

Appendix B: A Timetable of Partial Methods in Cell Type-Specific Analysis

Time and Journal	Title	Paper Link
Genome research, 01/2022, 32(01)	**Cell type–specific analysis by single-cell profiling identifies a stable mammalian tRNA–mRNA interface and increased translation efficiency in neurons**	**https://genome.cshlp.org/content/32/1/97**
Translational psychiatry, 04/2022, 12(01)	**Allele-specific analysis reveals exon- and cell-type-specific regulatory effects of Alzheimer's disease-associated genetic variants**	**https://www.proquest.com/docview/2651906726?accountid=13151&pq-origsite=summon&forcedol=true**
Nucleic acids research, 05/2022, 50(W1)	**WebCSEA: web-based cell-type-specific enrichment analysis of genes**	**https://academic.oup.com/nar/article/50/W1/W782/6591520**
Genome Biology, 06/2022, 23,(01)	**SCADIE: simultaneous estimation of cell type proportions and cell type-specific gene expressions using SCAD-based iterative estimating procedure**	**https://www.proquest.com/docview/2678211322?accountid=13151&pq-origsite=summon&forcedol=true**
Nucleic acids research, 06/2022, 50（10）	**CT-FOCS: a novel method for inferring cell type-specific enhancer–promoter maps**	**https://academic.oup.com/nar/advance-articles**
Genome biology, 07/2022, 23(1)	**DeCAF: a novel method to identify cell-type specific regulatory variants and their role in cancer risk**	**https://doaj.org/article/a2fd4c6f9fa543ae8b9a902944c5992c**
Gigascience, 04/2022, 11	**scMAPA: Identification of cell-type–specific alternative polyadenylation in complex tissues**	**https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac033/6576244**
Frontiers in oncology, 2022, 12	**Integrative Analysis Identifies Cell-Type-Specific Genes Within Tumor Microenvironment as Prognostic Indicators in Hepatocellular Carcinoma**	**https://doaj.org/article/6f610b45604044fd9e504c8e38b8e2a5**

Genome research, 10/2021, 31(10)	Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data	https://genome.cshlp.org/content/31/10/1807
Molecular neurobiology,10/2021, 59(01)	Sex-Stratified Single-Cell RNA-Seq Analysis Identifies Sex-Specific and Cell Type-Specific Transcriptional Responses in Alzheimer's Disease Across Two Brain Regions	https://link.springer.com/article/10.1007/s12035-021-02591-8
The Laryngoscope, 09/2021, 131(S5)	Cell Type–Specific Expression Analysis of the Inner Ear: A Technical Report	https://onlinelibrary.wiley.com/doi/full/10.1002/lary.28765
The FASEB journal, 05/2021, 35(5)	Cell-type specific analysis of physiological action of estrogen in mouse oviducts	https://faseb.onlinelibrary.wiley.com/doi/full/10.1096/fj.202002747R
Genome Biology, 03/2021, 22(1)	Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells	https://www.proquest.com/docview/2502905657?accountid=13151&pq-origsite=summon&forcedol=true
BMC bioinformatics, 03/2021, 22(01)	Nonlinear ridge regression improves cell-type-specific differential expression analysis	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7986289/
Science advances, 07/2021, 7(31)	Single-cell analyses unravel cell type-specific responses to a vitamin D analog in prostatic precancerous lesions	https://www.science.org/doi/10.1126/sciadv.abg5982
BMC research notes, 03/2021, 14(1)	SeqEnhDL: sequence-based classification of cell type-specific enhancers using deep learning models	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7980595/
Brief Bioinform .2021 Jan 18;22(1):	SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references	https://pubmed.ncbi.nlm.nih.gov/31925417/
Nature genetics, 01/2021, 53(01)	WAPL maintains a cohesin loading cycle to preserve cell-type-specific distal gene regulation	https://www.proquest.com/docview/2477275192?accountid=13151&pq-origsite=summon&forcedol=true
Methods in molecular biology (Clifton, N.J.), 01/2021	Single Cell Type Specific RNA Isolation and Gene Expression Analysis in Rice Using Laser Capture Microdissection (LCM)-Based Method	https://link.springer.com/protocol/10.1007/978-1-0716-1068-8_18

Frontiers in cellular neuroscience, 2020, 14	Cell Type-Specific Gene Network-Based Analysis Depicts the Heterogeneity of Autism Spectrum Disorder	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7096557/
Nature communications, 02/2020, 11, 1	RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7039879/
PloS one, 2020, 15(4)	SMART-Q: An Integrative Pipeline Quantifying Cell Type-Specific RNA Transcription	
PLoS genetics, 04/2020, 16(04)	An integrated analysis of cell-type specific gene expression reveals genes regulated by REVOLUTA and KANADI1 in the Arabidopsis shoot apical meristem	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7266345/
Neuroscience research, 03/2020, 152	Enhancer-Driven Gene Expression (EDGE) enables the generation of cell type specific tools for the analysis of neural circuits	https://www.sciencedirect.com/science/article/pii/S0168010220300328?via%3Dihub
Bioinformatics, Volume 36, Issue 4, 15 February 2020,	scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data	https://academic.oup.com/bioinformatics/article/36/4/1262/5564118
Bioinformatics, 02/2020, 36(03)	Using multiple measurements of tissue to estimate subject- and cell-type-specific gene expression	https://academic.oup.com/bioinformatics/article/36/3/782/5545976
Science signaling, 02/2020,13（620）	Integrative analysis suggests cell type-specific decoding of NF-κB dynamics	https://www.science.org/doi/10.1126/scisignal.aax7195
PloS one, 2020, 5(4)	SMART-Q: An Integrative Pipeline Quantifying Cell Type-Specific RNA Transcription	https://www.researchgate.net/publication/338827850_SMART-Q_An_Integrative_Pipeline_Quantifying_Cell_Type-Specific_RNA_Transcription
Genome Biology, 08/2020, 21(1)	3DeFDR: statistical methods for identifying cell type-specific looping interactions in 5C and Hi-C data	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7496221/
Journal of proteome research, 01/2020, 19(01)	Extended Human G-Protein Coupled Receptor Network: Cell-Type-Specific Analysis of G-Protein Coupled Receptor Signaling Pathways	https://pubs.acs.org/doi/10.1021/acs.jproteome.9b00754

Cell systems, 12/2019, 9(6)	Reconstruction of Cell-type-Specific Interactomes at Single-Cell Resolution	https://www.sciencedirect.com/science/article/pii/S2405471219303837?via%3Dihub
Nucleic Acids Research, Volume 47, Issue 19	Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data	https://academic.oup.com/nar/article/47/19/10027/5566587
Bioinformatics.2019 Oct 15;35(20)	Dissecting differential signals in high-throughput data from complex tissues	https://pubmed.ncbi.nlm.nih.gov/30903684/
Methods (San Diego, Calif.), 08/2019, 166	FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data	https://www.sciencedirect.com/science/article/pii/S1046202318303293?via%3Dihub
Nature protocols, 08/2019, 14(8)	Sequencing cell-type-specific transcriptomes with SLAM-ITseq	https://www.proquest.com/docview/2564692180?pq-origsite=summon
Nature Biotechnology Pub Date : 2019-05-06	Determining cell type abundance and expression from bulk tissues with digital cytometry.	https://www.x-mol.com/paper/5673877
Nature Communications volume 10,	Accurate estimation of cell-type composition from gene expression data	https://www.nature.com/articles/s41467-019-10802-z
Nucleic acids research, 11/2019, 47(19)	Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6821429/
Nature Communications 2019.volume 10,	Bulk tissue cell type deconvolution with multi-subject single-cell expression reference	https://www.nature.com/articles/s41467-018-08023-x
BMC genomics, 06/2018, 19(1)	Cell type-specific analysis of transcriptome changes in the porcine endometrium on Day 12 of pregnancy	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6000939/
Cell Systems Volume 6, Issue 4, 25 April 2018,	Systematic Evaluation of Molecular Networks for Discovery of Disease Genes	https://www.sciencedirect.com/science/article/pii/S2405471218300954
Development	SLAM-ITseq: Sequencing cell type-specific transcriptomes	https://journals.biologists.com/dev/article/145/13/dev164640/19298/SLAM-ITseq-sequencing-

Cambridge), 01/2018, 145(13)	**without cell sorting**	**cell-type-specific**
Biological psychiatry (1969), 2016, 81(3)	**A Comprehensive Analysis of Cell Type–Specific Nuclear RNA From Neurons and Glia of the Brain**	**https://pku.summon.serialssolutions.com/search?s.q=cell-type-specific#!/search?ho=t&include.ft.matches=f&l=zh-CN&q=cell-type-specific%20analysis**
Nature protocols, 09/2016, 11(9)	**Cell-type-specific profiling of protein-DNA interactions without cell isolation using targeted DamID with next-generation sequencing**	**https://www.proquest.com/docview/1810123264?pq-origsite=summon&parentSessionId=x2ciW%2Fyr%2BWF10BoyOVip0maX4buU%2FVDzAXGWlsVvuvI%3D**
BMC bioinformatics, 12/2015, 16, 413	MixChIP: a probabilistic method for cell type specific protein-DNA binding analysis	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4690251/
BMB reports, 07/2015, 48(7)	**Cell type-specific gene expression profiling in brain tissue: comparison between TRAP, LCM and RNA-seq**	**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4577288/**
Nature Methods 2015 May;12(5)	**Robust enumeration of cell subsets from tissue expression profiles**	**https://pubmed.ncbi.nlm.nih.gov/25822800/**
BMC bioinformatics, 07/2015, 16(1)	**groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data**	**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4502638/**
Hox Genes, 07/2014	**cgChIP: A Cell Type- and Gene-Specific Method for Chromatin Analysis**	https://link.springer.com/protocol/10.1007/978-1-4939-1242-1_18
Bioinformatics, 03/2014, 30(5)	**MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples**	https://academic.oup.com/bioinformatics/article/30/5/682/244931
PLoS Comput Biol. 2014;10	**Enhanced regulatory sequence prediction using gapped k-mer features.**	https://pubmed.ncbi.nlm.nih.gov/25033408/
Genome research, 05/2014, 24(5)	**General approach for in vivo recovery of cell type-specific effector gene sets**	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4009615/
PLoS genetics, 2013, 9(10)	**Genome-Wide Analysis of Cell Type-Specific Gene Transcription during Spore Formation in Clostridium**	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3789822/

	difficile	
BMC bioinformatics, 03/2013, 14(1)	**Digital sorting of complex tissues for cell type-specific gene expression profiles**	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3626856/
Genome research, 09/2012, 22(09)	**Predicting cell-type-specific gene expression from regions of open chromatin**	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431488/
BMC genomics, 11/2012, 13(1)	Cell population-specific expression analysis of human cerebellum	https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-610
Genome Res. 2011;21	Discriminative prediction of mammalian enhancers from DNA sequence.	https://pubmed.ncbi.nlm.nih.gov/21875935/
Bioinformatics, Volume 27, Issue 6, 15 March 2011	Identifying dispersed epigenomic domains from ChIP-Seq data	https://academic.oup.com/bioinformatics/article/27/6/870/236489
Nature protocols, 01/2011, 6(01)	The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana	https://www.proquest.com/docview/1041013576?pq-origsite=summon
Molecular systems biology, 10/2010, 6(01)	**Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control**	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2990639/
Nature biotechnology, 05/2010, 28(5)	**Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs**	https://www.proquest.com/docview/222294131?accountid=13151&pq-origsite=summon&forcedol=true
Molecular Cell Volume 38, Issue 4, 28 May 2010,	**Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities**	https://www.sciencedirect.com/science/article/pii/S1097276510003667
Nature methods, 04/2010, 7(04)	**Cell type–specific gene expression differences in complex tissues**	https://www.nature.com/articles/nmeth.1439#Sec2
Bioinformatics, Volume 25, Issue 15	**A clustering approach for identification of enriched domains from histone modification ChIP-Seq data**	https://academic.oup.com/bioinformatics/article/25/15/1952/212783
BMC genomics	**Correlation of mRNA and protein levels: Cell type-**	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2413246/

, 05/2008, 9(01)	**specific gene expression of cluster designation antigens in the prostate**	