Distribution Agreement

Signature:

_________________________________ ________________

Siyi Geng Date

Cell Type–specific Gene Expression and DNA Methylation Differences in Complex Tissues

By

Siyi Geng

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

____

Hao Wu, Ph.D.

Committee Chair

____

Xiangqin Cui, Ph.D.

Committee Member

Cell Type–specific Gene Expression and DNA Methylation Differences in Complex Tissues

By

Siyi Geng

B.E.

Nanjing Agricultural University

2016

Thesis Committee Chair: Hao Wu, Ph.D.

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics

2019

Abstract

Cell Type–specific Gene Expression and DNA Methylation Differences in Complex Tissues

By Siyi Geng

A majority of tissues such as blood and tumor are complex and heterogeneous samples containing different cell types. Thus, the profiles of the genome or epigenome of tissue samples from high-throughput technologies are mixed signals. The heterogeneity in such data bring difficulties in data analysis and result in biases without proper adjustment.

We extend an existing method TOAST (TOols for the Analysis of heterogeneouS Tissues) to model the data from mixed, heterogeneous samples and detect cell-specific differential signals. We design a series of simulation studies on cell-specific differential expression (csDE) detection to evaluate the TOAST performance. Furthermore, we conduct analysis on DNA methylation (DNAm) data from two existing human blood datasets. We use a reference-based method EpiDISH to estimate cell proportions and apply TOAST to detect age-related cell-specific differential methylated CpG sites (csDMC).

Simulation studies and analysis on real data show good performance of upgraded TOAST on csDE/DM detection. The results from the simulation study show that larger sample size has a positive effect on performance accuracy, while the larger noise level has a negative effect. In real data study, we find that age is related to cell proportions of mixed samples. Through csDM analysis using TOAST, we identify varies of age-related DMCs in each cell type and the numbers of csDMCs are different among cell types. These results show that the upgraded TOAST provides a flexible statistical method to analyze cell-specific differential gene expression and DNA methylation.

Cell Type–specific Gene Expression and DNA Methylation Differences in Complex Tissues

By

Siyi Geng

B.E.

Nanjing Agricultural University

2016

Thesis Committee Chair: Hao Wu, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics

2019

Acknowledgment

I would first like to thank my thesis advisor Dr. Hao Wu, Associate Professor at our Department of Biostatistics and Bioinformatics at Emory University. His guidance helped me in all the time of research and writing of this thesis. I really enjoyed working with him and learned a lot from him not only academically but also personally.

I would also express great thanks to my senior Dr. Ziyi Li, Postdoc at Emory University. Her patience and enthusiasm have supported me to complete my thesis. She has been my mentor and friend. It has been my pleasure working with her. I am also grateful to my thesis reader Dr. Cui for her help during the revision process and the guidance of the entire essay writing.

I would like to extend my thanks to my beloved friend Jingyi Liu for teasing the difficulties and life with me and making my life full of humor. Finally, I would like to thank my family for their continuous support and encouragement. I thank my parents for always encouraging me to pursue being who I want to be.

Contents

Introduction

Background of Genetics and Genomics

Gene is a segment of DNA that codes functional molecules, which is the basic physical and functional unit of heredity. Gene expression is the process by which the genetic information in the DNA is implemented into functional products such as proteins or other molecules. For synthesizing proteins, gene expression has two key steps: transcription and translation, both are tightly regulated by many mechanisms.

DNA methylation (DNAm) is an epigenetic process where a methyl group is added to the DNA molecule. In mammals, more than 98% DNA methylation occurs in a CpG dinucleotide context in somatic cells (Jin, Li, and Robertson 2011). Methylation often affects gene expressions and functions. Existing studies reported that cytosine methylation is widespread in animals but the patterns vary in space and time (Bird 2002). In mammals, 70% to 80% of all cytosines in CpG site (the dinucleotide where C and G appear at consecutive bases) are methylated (Jabbari and Bernardi 2004). A set of special genomic regions is the CpG island, which has high frequency of CpG sites and is usually associated with promoter regions. Generally, CpG islands maintain an unmethylated state (Christensen et al. 2009). An increasing body of studies reports association between aging and epigenetic mechanism, particularly, age-related DNA methylation (Horvath 2016; Gregory Hannum et al. 2013; Christensen et al. 2009; Jaffe and Irizarry 2014).

Knowledge of DNA sequences has greatly enhanced our understanding for biological processes and become essential in both basic biological research and clinical application. DNA sequencing is the technology of determining the sequence of the nucleotides in DNA.

It is usually accomplished by cutting long DNA fragments (like chromosomes) into short pieces, creating multiple copies for each piece through PCR, and sequencing all of them through one of several technologies (Mäkinen et al. 2015). “High-throughput sequencing” has revolutionized the genomics research since its invention in early 2000’s. Compared to earlier low through-put DNA sequencing methods, the high-throughput methods can sequence much larger amount of DNA in a short period time, and with much lower cost.

Sample Mixture Problem

In a practical research environment, people often profile the genome or epigenome of tissue samples using high-throughput technologies. The tissue samples, such as blood, brain, and tumor, are mostly very heterogeneous. Tumor tissue, as an example, contains distinct cell types, as demonstrated by in vivo multilineage differentiation and clonal genetic heterogeneity (Shipitsin et al. 2007; Dalerba et al. 2011). It was shown that tumor samples are mixtures of tumor cells, precancerous lesion cells, and surrounding normal cells. The variation in gene expression and DNA methylation from distinct cell and tissue types has been reported from multiple studies (Walker et al. 1983; Whitney et al. 2003; Shen-Orr et al. 2010; Heintzman et al. 2009; Brady et al. 1995; Absher et al. 2013; Ho et al. 1993). Because of this, the overall gene expression or DNA methylation measurements from complex tissues are often affected by the cell mixtures. In human blood, the pattern of variation in gene expression is correlated with certain cell counts: CD20 gene expression associated with lymphocyte count, while the variations of CD19, CD22 and CD72 genes expressions are affected by B cells (Whitney et al. 2003). In a systemic lupus erythematosus study testing the differential DNA methylation (DM) between patients and controls, the patterns of methylation difference varied among T cells, B cells and monocytes with different methylated CpGs (Absher et al. 2013). Thus, the heterogeneous and dynamic nature of complex tissues (Clarke et al. 2008) is a confounding factor for data analysis and requires further investigation.

The Existing Statistical Methods

High-throughput measurements from complex tissues containing different cell types are mixed signals (weighted averages of the pure cell type quantities). Traditional analysis methods for differential expression (DE) and differential methylation (DM) often ignore this problem, which could lead to biased result. Recently, the method development for high-throughput data from complex tissues has gained much interest (Shen-Orr and Gaujoux 2013; Brady et al. 1995; Achim et al. 2015; Whitney et al. 2003). In cell type specific research, cell isolation methods, such as Fluorescence-activated cell sorting (FACS) and Microfluidic, are widely applied. However, these methods are expensive and require high technical skills (Hu et al. 2016). In addition, these methods are limited by the instability of protein epitopes, the requirements for cell processing and the timeliness of cell analysis, which is beyond the possibility to analyze the normally mixed composition of tissues (Houseman et al. 2012). In multiple studies, scientists estimated cell proportion through global profiles (Teschendorff et al. 2017; Zheng et al. 2018). *EpiDISH*, as an example, generate estimated cell proportions based on robust partial correlations (Teschendorff et al. 2017).

The development of computational methods for analyzing csDE/csDM from heterogeneous tissues began almost two decades ago, and gained significant interest recently. (Venet et al. 2001; Erkkilä et al. 2010). In an early work (Venet et al. 2001), a linear model was proposed, assuming that each cell type has some uniquely expressed genes as cell type markers, by which both the cell-type specific frequency and cell-type-specific gene expression are estimated. Population-Specific Expression Analysis (PSEA) assumes the cell type specific gene expression to be proportional to the cell proportions and builds a linear regression model for cell-specific marker genes (Kuhn et al. 2011). The application of these methods has several requirements and lacks flexibility. PSEA, as an example, relies on marker genes expression, which is only available for tissues with identified cell-specific marker genes (Kuhn et al. 2012).

Age Impact on DNA Methylation

According to multiple studies, age-related DNAm changes have been identified in multiple tissues and organisms. In human whole blood, genome-scale DNAm profiling identifies many CpGs with methylation levels significantly associated with age (Rakyan et al. 2010). Changes in methylation have been linked to complex age-associated phenotypes such as telomere length and systolic blood pressure (SBP) (Bell et al. 2012), age-associated diseases such as Alzheimer's Disease (Wang, Oeize, and Schumacher 2008), and cancer (Levine et al. 2015). The association between age and DNAm is also modeled to measure and compare human aging rates (Gregory Hannum et al. 2013).

Study Goals

The goals in this thesis work are two-fold. First, we will extend the previously developed method TOAST (TOols for the Analysis of heterogeneouS Tissues), which detects differential signals in high-throughput data from mixed samples. TOAST has been implemented as an R package and is freely available on GitHub (https://github.com/ziyili20/TOAST). Although the original TOAST paper proposed a generalized framework for identifying cell-type specific differential signals from various conditions, the implementation only considers two-group design, i.e. samples from diseased and controls. In this work, we will expand the implementation to more general experimental designs, which accepts various covariates including binary, categorical, and continuous variables. We will examine our implementation through simulation studies and evaluate the impact of sample size and noise levels. Secondly, we will apply the method on two real data sets to identify age-related cell-specific differential methylated CpG sites (DMCs).

Method

Estimating mixture proportions using EpiDISH

With high-throughput measurements of mixed samples, the first step of the analysis is to estimate mixture proportions for different cell types. In this study, we use an existing reference-based deconvolution method *EpiDISH* to estimate the cell proportions (Teschendorff et al. 2017; Houseman et al. 2012; Newman et al. 2015).

EpiDISH models the observation of mixed sample from one subject $\boldsymbol{Y}$ as a linear combination of *K* cell-type specific pure profiles $\boldsymbol{h_k}$

$$\mathbf{Y} = \sum\nolimits_{k=1}^{K} w_k \boldsymbol{h_k} + \varepsilon$$

Here w_k denotes the weight coefficients of the k^{th}cell type. satisfying a constraint $\sum_{k=1}^{K} w_k = 1$. EpiDISH uses robust partial correlations (RPC) to estimate w_k and enforce the constraints a *posteriori* by setting all w_k non-negative and added up to 1, following the procedures described in Newman et al. 2015. *Lm()* and *rlm()* functions in R are used to perform multivariate regressions for RPC in EpiDISH. EpiDISH is available as an R-package from Bioconductor: https://bioconductor.org/packages/release/bioc/html/EpiDISH.html.

TOAST Model

Assuming we have high-throughput measurements **Y** for **N** samples and **G** features (e.g. genes, CpG sites, ... etc.), and each subject has **K** underlying "pure" cell types, the mixing proportions $\boldsymbol{\theta_i} = (\theta_{i1}, \theta_{i2}, \dots, \theta_{ik})$ for i^{th} sample (under the constraint $\sum_1^K \theta_{ik} = 1$) can be estimated using the above described method EpiDISH. $\mathrm{Y_{gi}}$ denotes the measurement for the $\mathrm{g^{th}}$ feature in the $\mathrm{i^{th}}$ sample. $\mathrm{X_{gik}}$ denotes the measurement for the $\mathrm{g^{th}}$ feature in the $\mathrm{i^{th}}$ sample in the $\mathrm{k^{th}}$ cell type.

Denote the vector for subject-specific covariates by $\mathbf{Z_i}$ for $\mathrm{i^{th}}$ sample. In the most general case, $\mathbf{Z_i}$ can contain a number of covariates that are binary, categorical, or continuous. All covariates are encoded in a typical fashion in linear regression. For example, categorical covariates can be represented as a vector of dummy variables. For example, a three-level covariate has two degrees of freedoms and will be coded as $(0,0)$ for Group 1 (reference), $(1,0)$ for Group 2, and $(0,1)$ for Group 3. A continuous variable will have one degree of freedom. Overall, assume there are $\boldsymbol{M}$ categorical variables, and the $\mathrm{m^{th}}$ covariate contains υ_m levels, and $\boldsymbol{P}$ continuous variables in the model, the number of elements in vector $\mathbf{Z_i}$ is

$$\sum_m (\upsilon_m - 1) + \boldsymbol{P}.$$

We can write the profile for the kth cell type as: $\mathrm{E}[\mathrm{X}_{\mathrm{ik}}] = \mu_{\mathrm{k}} + \mathbf{Z_i}^{\mathbf{T}}\boldsymbol{\beta}_{\mathbf{k}}$, where μ_{k} is the baseline for kth cell type, and $\boldsymbol{\beta}_{\mathbf{k}}$ are defined as the covariate's coefficients for the kth cell type. Usually, X_{ik} is not directly observed in experiments. Instead, the weighted average of X_{ik} by mixing proportions $\boldsymbol{\theta}_{\boldsymbol{i}}$ is observed. Then, the ith sample with given $\boldsymbol{\theta}_{\boldsymbol{i}}$ will have:

$$\mathrm{E}[\mathrm{Y}_{\mathrm{i}}; \boldsymbol{\theta}_{\mathbf{i}}] = \sum_{k} \theta_{ik} E[X_{ik}] = \sum_{k} (\theta_{ik}\mu_{\mathrm{k}} + \theta_{ik}\mathbf{Z_i}^{\mathbf{T}}\boldsymbol{\beta}_{\mathbf{k}})$$

This is a linear model with unknown parameters $\boldsymbol{\mu_k}$ *and* $\boldsymbol{\beta_k}$. The mixing proportions $\boldsymbol{\theta_i}$ and mixing proportions by covariate interactions $\boldsymbol{\theta_i Z_i}$ are known and included in the design matrix. Thus, we can write the model in a matrix form as

$$\mathrm{E}[\mathbf{Y}] = \mathbf{W}\boldsymbol{\beta}$$

$$\textbf{where } \mathbf{W} = \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1K} & \theta_{11} \cdot Z_1{}^T & \theta_{12} \cdot Z_1{}^T & \dots & \theta_{1K} \cdot Z_1{}^T \\ \theta_{21} & \theta_{22} & \dots & \theta_{2K} & \theta_{21} \cdot Z_2{}^T & \theta_{22} \cdot Z_2{}^T & \dots & \theta_{2K} \cdot Z_2{}^T \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{N1} & \theta_{N2} & \dots & \theta_{NK} & \theta_{N1} \cdot Z_N{}^T & \theta_{N2} \cdot Z_N{}^T & \dots & \theta_{NK} \cdot Z_N{}^T \end{bmatrix},$$

$$\boldsymbol{\beta} = [\mu_1 , \mu_2 , \dots , \mu_k , \boldsymbol{\beta_1^T}, \boldsymbol{\beta_2^T}, \dots , \boldsymbol{\beta_k^T}]^T .$$

$\boldsymbol{B}$ has $\mathrm{Q} = \mathrm{K} + \sum_m (\upsilon_m - 1) + \boldsymbol{P}$ elements.

Simulation

We design a series of simulation studies to evaluate our implementation of the TOAST methods. All simulations are focused on csDE detection, since there are good cell type specific expression data that can be used for simulation. In all simulations, model parameters applied to generate data are estimated from real data in order to obtain a realistic simulation setting.

First, cell-type specific profiles are generated based on a real dataset studying immune system (Wolslegel et al. 2009). The data were generated in a systemic lupus erythematosus study, which includes the gene expression profiles from four types of peripheral leukocytes cell lines (B cells: "Raji" and "IM-9"; T cell: "Jurkat"; monocyte cell: "THP-1"). Based on this dataset, the mean μ_{gk} and variance of log expression values were estimated, for the gene in the cell type. Then we obtained the estimated profiles $\mathbf{X}_{\mathrm{i}}$ with 54657 genes and 4 cell lines for the i^{th} subject from a log-normal distribution with mean $(\mu_{gk})_{GK}$ and variance $({\sigma^2}_{gk})_{GK}$.

Next, we simulate three subject-level covariates, including a binary one (such as disease, 1 represents case, and 0 for control), a continuous one (such as age, 30 - 50), and a categorical one with 3 groups (such as ethnicity). We simulate the data this way in order to test the functionality of TOAST to make sure it works for all type of variables (continuous, categorical with 2 or more levels). We define a 30-year-old normal subject in ethnicity 1 as a baseline subject. For baseline subjects, we simulate the $(\mathrm{g}^{\mathrm{th}}, \mathrm{k}^{\mathrm{th}})$ element of $\mathbf{X}_{\boldsymbol{i}}$ from a log-normal distribution with mean μ_{gk} and variance ${\sigma^2}_{gk}$. For other subjects, we randomly select a group of genes in each cell types as csDE genes related with different phenotypes, half of them are up-regulated and another half are down-regulated. To be specific, 3% DE genes are selected for case subjects versus controls, similarly, the other two groups of 3% genes are selected independently for ethnicity 2 and ethnicity 3 subjects versus ethnicity 1 subjects. The log fold changes (lfc) are randomly drawn from N(±1, 0.2) for up- and down-regulated genes. Age are generated randomly from uniform distribution Unif(30, 50). The lfc per unit age increase are randomly applied from N(±0.03, 0.1) for up- and down-regulated genes. For example, the i^{th} subject, who is a 30-year-old patients in ethnicity 1, D = [0.015*G] ([.] is the round operation) genes are selected to be up-regulated and another D genes are down-regulated if this is a patient. Then we have:

$$\boldsymbol{X}_{ik} \sim \log - normal \left(\begin{bmatrix} \mu_{1k} \\ \vdots \\ \mu_{Dk} \\ \mu_{(D+1)k} \\ \vdots \\ \mu_{(2D)k} \\ \mu_{(2D+1)k} \\ \vdots \\ \mu_{Gk} \end{bmatrix} + \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \times N(1, 0.2), \begin{bmatrix} {\sigma^2}_{1k} \\ \vdots \\ {\sigma^2}_{Dk} \\ {\sigma^2}_{(D+1)k} \\ \vdots \\ {\sigma^2}_{(2D)k} \\ {\sigma^2}_{(2D+1)k} \\ \vdots \\ {\sigma^2}_{Gk} \end{bmatrix} \right)$$

X_{ik} denotes the underlying expression level for the G genes in the i^{th} sample in the k^{th} cell type (here the subscript for gene is omitted for notation simplicity). We consider 50, 100, and 500 as total sample sizes, and in all settings the total sample size are distributed equally by disease status (case vs control) and ethnicity levels (1, 2 and 3).

Furthermore, the mixing proportions $\boldsymbol{\theta}_{\boldsymbol{i}}$ are generated based on the estimated proportions from an Alzheimer's study dataset (Sonnen et al. 2009). We estimated the MLE of $\boldsymbol{\alpha}_{\mathbf{0}}$ = (0.968, 4.71, 0.496, 0.347) based on the four cell proportions of 11 normal subjects in Sonnen et al. study. Dirichlet distribution $\boldsymbol{\theta}_{\boldsymbol{i}} \sim \mathbf{Dir}(\boldsymbol{\alpha}_{\mathbf{0}})$ is applied to generate proportions $\boldsymbol{\theta}_{\boldsymbol{i}}$.

Lastly, with the simulated $\mathbf{X}_{\boldsymbol{i}}$ and proportions $\boldsymbol{\theta}_{\boldsymbol{i}}$**,** the measurement for the i^{th} subject is obtained by $\mathbf{Y_i} = \boldsymbol{\theta}_{\boldsymbol{i}}\mathbf{X_i} + \boldsymbol{\varepsilon}$**.** $\boldsymbol{\varepsilon}$ is the measurement error, and randomly generated from $\mathrm{N}(0, n_{sd}\eta_g^2)$ based on the Immune data. n_{sd} is the technical noise level, three levels are selected in simulation study: $n_{sd} = 0.1$ for low level, 1 for medium level and 10 for high level. η_g is the standard deviation of measurement error in the g^{th} gene. η_g is simulated based on a function of $\overline{\mathrm{X}_{\mathrm{i}}\theta_{\mathrm{i}}} = \frac{1}{K}\sum_{k}\mathrm{X}_{\mathrm{gik}}\theta_{\mathrm{ik}}$, $\eta_{g} = -8.06 + 0.11\overline{\mathrm{X}_{\mathrm{i}}\theta_{\mathrm{i}}}$, which is estimated from the Immune Dataset.

Real Data Application

To test the functionality of TOAST on detecting cell-type specific DNAm changes, we applied our method on two real DNA methylation datasets from human blood tissues. The data were obtained from NCBI GEO database, both datasets were generated from Illumina Infinium HumanMethylation450 BeadChip.

The first dataset (GSE42861) is from a study aiming to detect the methylation differences from 354 Rheumatoid arthritis patients (cases) and 335 normal controls using their peripheral blood leukocytes (PBLs) (Liu et al. 2013). In this study, we only use the 335 normal subjects as Rheumatoid arthritis may cause cell type changes in patients' blood. The second dataset (GSE40279) is a DNA methylation dataset measured from 656 individuals' whole blood samples. The original study was designed to characterize the association between the genome-wide methylation and human aging rates (Gregory Hannum et al. 2013).

In this project, we assume both datasets, either gathered from peripheral blood leukocytes or whole blood samples, contain six types of blood cells: CD8+ T cell, CD4+ T cell, natural killer cells (NK), B cells, monocytes (Mono), and granulocytes (Gran). We first use EpiDISH to estimate the proportions. For that, the top 10000 cell marker CpGs are selected through TOAST based on the sorted specificity index and are imported into EpiDISH to conduct deconvolution for estimating the cell proportions with RPC method.

Next, the cell-specific differential methylated CpG sites (DMCs) among different age are detected through TOAST based on the mixed sample profiles and estimated cell proportions. We test the cell-specific coefficients for age impact and calculate the Benjamini-Hochberg False Discovery Rate (FDR). The true DMCs are defined as FDR smaller than 0.05. Finally, we explore the overlapped DMCs for each of six cell types between two datasets and conduct gene ontology (GO) analysis through online gene enrichment analysis tool Enrichr.

Enrichment Analysis

Gene set enrichment analysis tool *Enrichr* (Chen et al. 2013; Kuleshov et al. 2016) (Online tool available at http://amp.pharm.mssm.edu/Enrichr/) is applied for gene ontology (GO) enrichment analysis. First, each DMC is mapped to its nearest gene based on gene annotation for probes on Illumina Infinium methylation 450k methylation microarrays of human genome version hg19. An R-package *IlluminaHumanMethylation450kanno.ilmn12.hg19* from Bioconductor is used for that (http://bioconductor.org/packages/IlluminaHumanMethylation450kanno.ilmn12.hg19/). After importing overlapped DMCs list, it returns a list containing gene location information. In Enrichr, we choose hg19 dataset as reference libraries and set 1000 genes as the maximum entry number. Upload gene list into Enrichr, it returns a result table in which each imported gene is associated with a functional term or an enrichment term including p-values, adjusted p-values (FDR), z-scores and combined scores. The p-values are calculated by Fisher's exact test to examine the binomial distribution assumption of DMC-mapped genes and the independence for the probability of any gene set containing certain gene. The adjusted p-values (FDR) are computed based p-values using the Benjamini-

Hochberg method for multiple testing correction. The rank scores or z-scores measure the deviation from the expected rank and are computed through a modification to Fisher's exact test. Finally, the combined scores are equal to the natural logarithm of p-values multiplied by z scores. We choose combined scores for enrichment terms ranking since it combines both methods.

Results

Simulation

In the simulation study, we evaluated the method to detect cell-specific csDE from microarray data. We conducted simulations in three ways: 1) age effect comparison; 2) case vs control comparison; and 3) ethnicity comparison: ethnicity 2 vs ethnicity 1, and ethnicity 3 vs ethnicity 1. The impact of sample size and noise levels on the accuracy of csDE detection were evaluated. The True Discovery Rate (TDR, which is the percentage of true positives) among top-ranked genes was calculated to measure the method performance. Each simulation was run 20 times, and the results are the average of the 20 iterations. The typical setting in simulation studies are with medium sample size 100 and

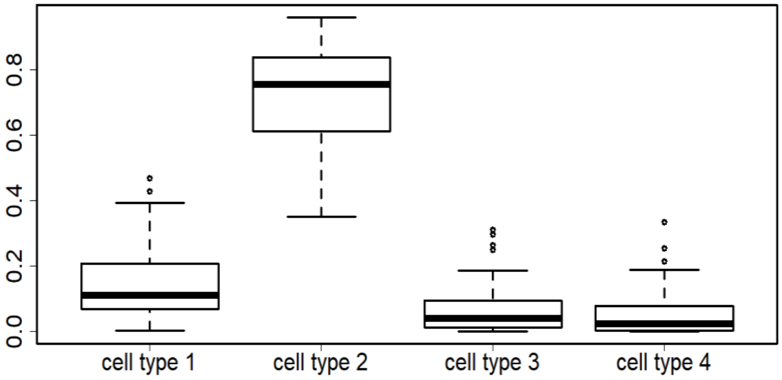


Figure 1 Boxplot of the estimated proportions

medium noise level. We assumed there are four cell types in the mixed sample, and the cell proportions were estimated under the reference-based method as mentioned in Method (Figure 1). The data contains three training groups with modest sample size (total 100 samples), medium noise level ($n_{sd} = 1$, which is the baseline noise level estimated from the Immune Data) and the age range from 30 to 50.

Impact of noise level and sample size

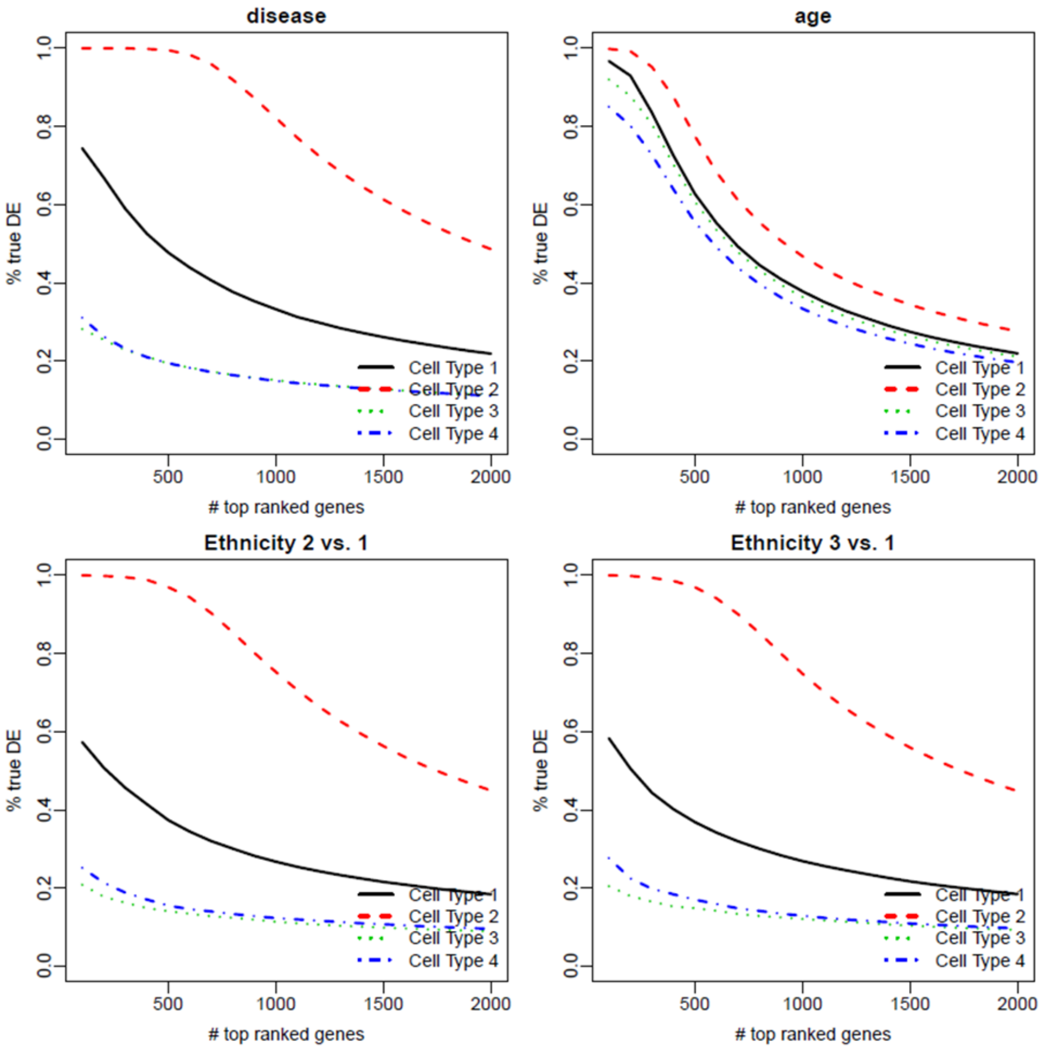


Figure 2a Impact of sample size on DE detection accuracy – Sample Size = 50

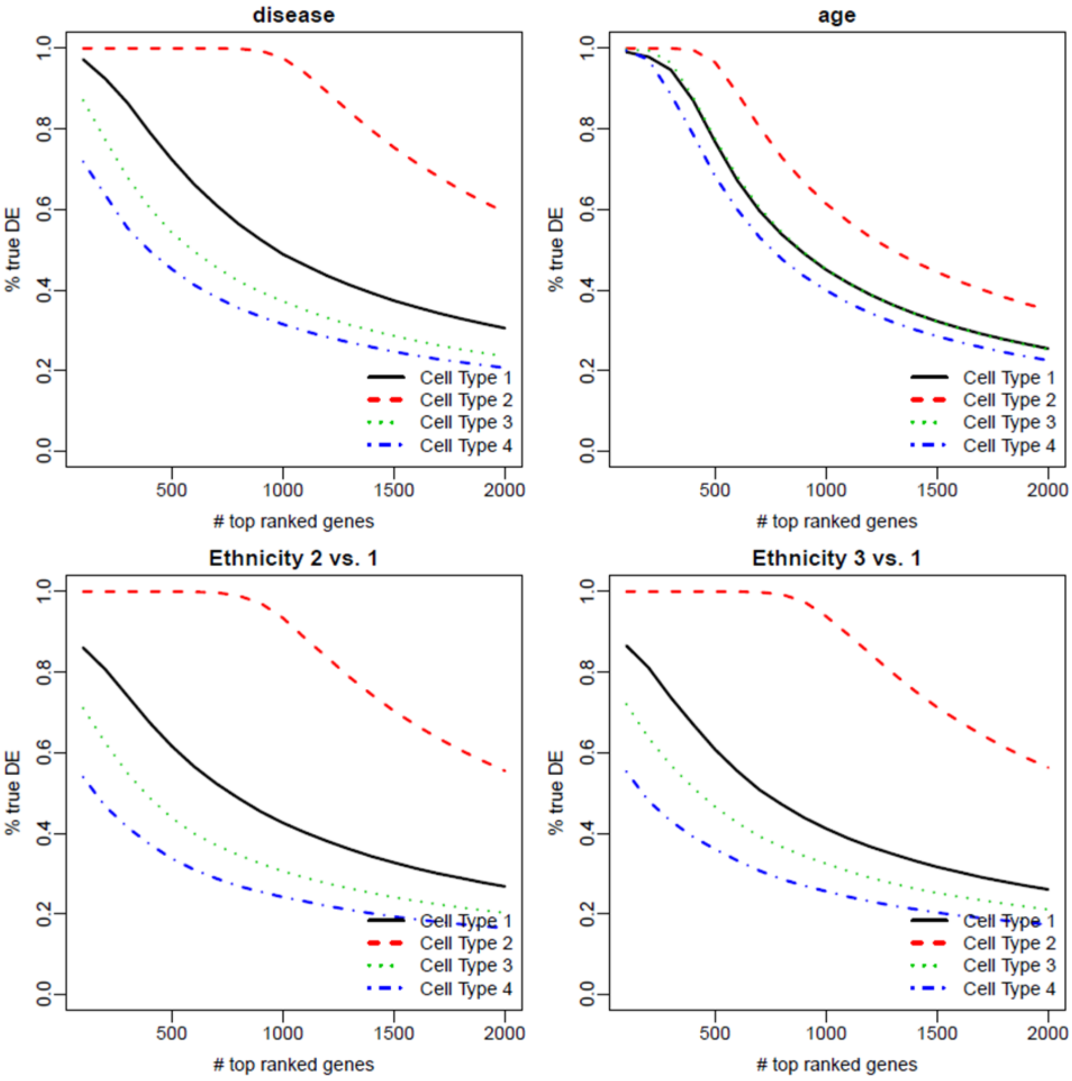


Figure 2b Impact of sample size on DE detection accuracy – Sample Size = 100

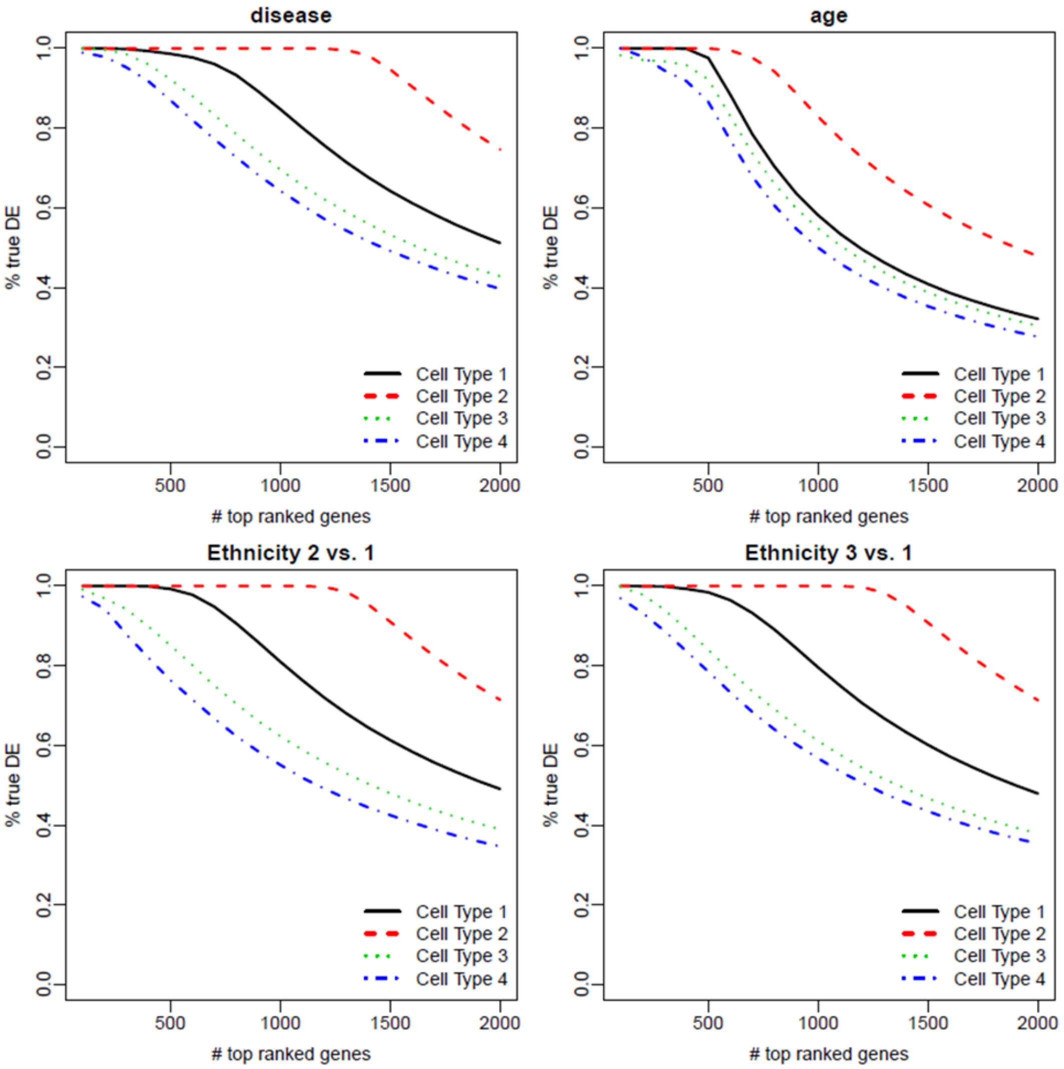


Figure 2c Impact of sample size on DE detection accuracy – Sample Size = 500

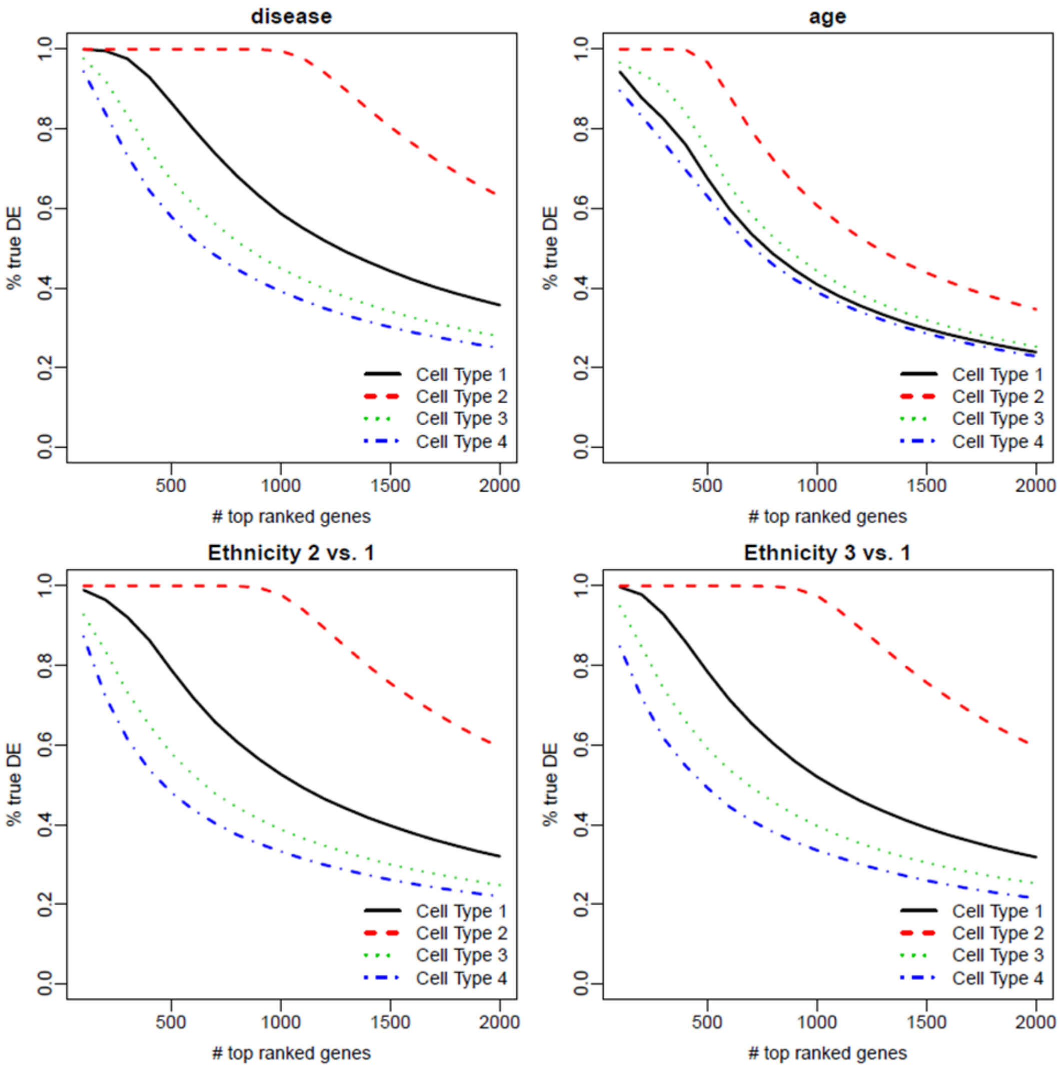


Figure 3a Figure 2c Impact of noise level on DE detection accuracy – n_sd=0.1

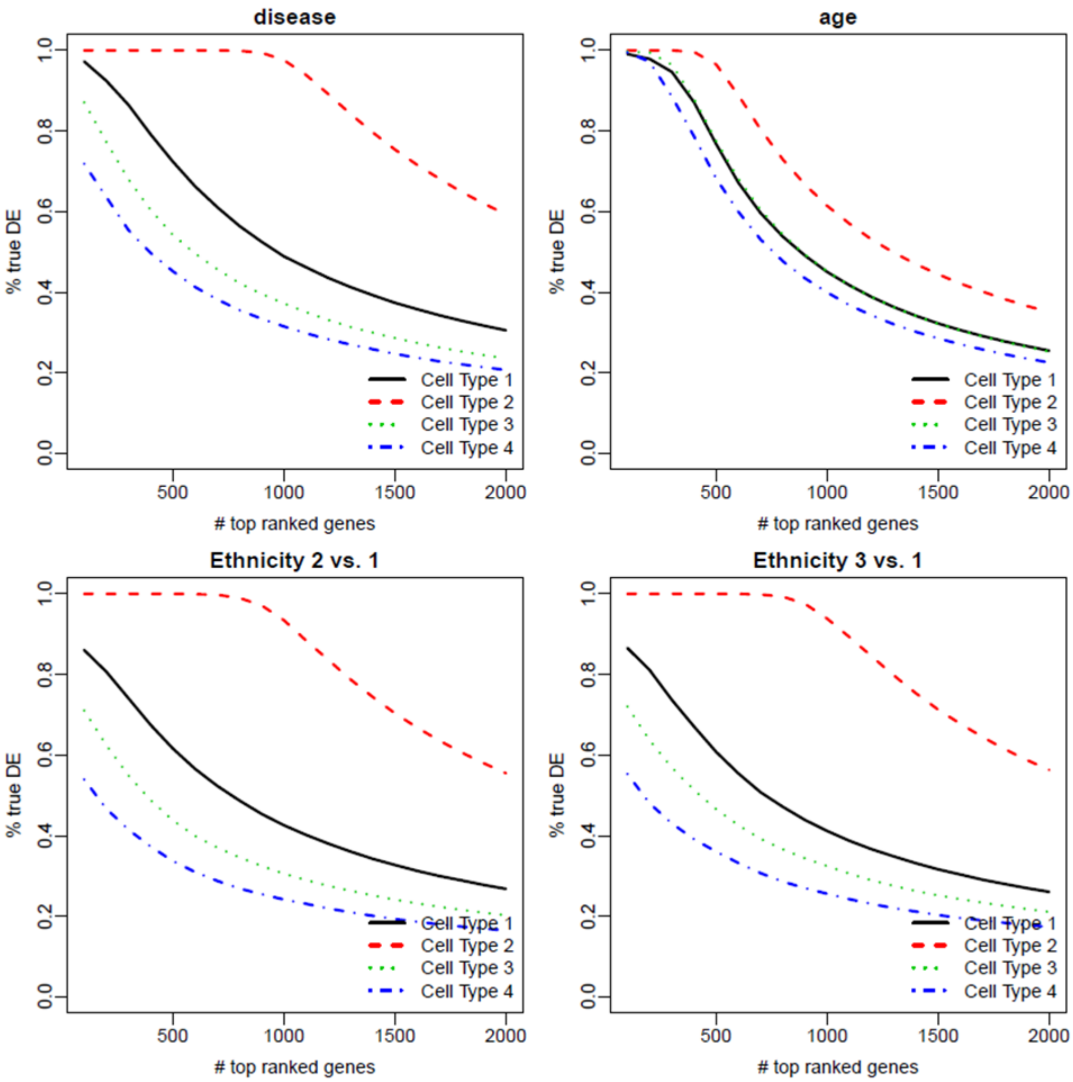


Figure 3b Impact of noise level on DE detection accuracy – n_sd = 1

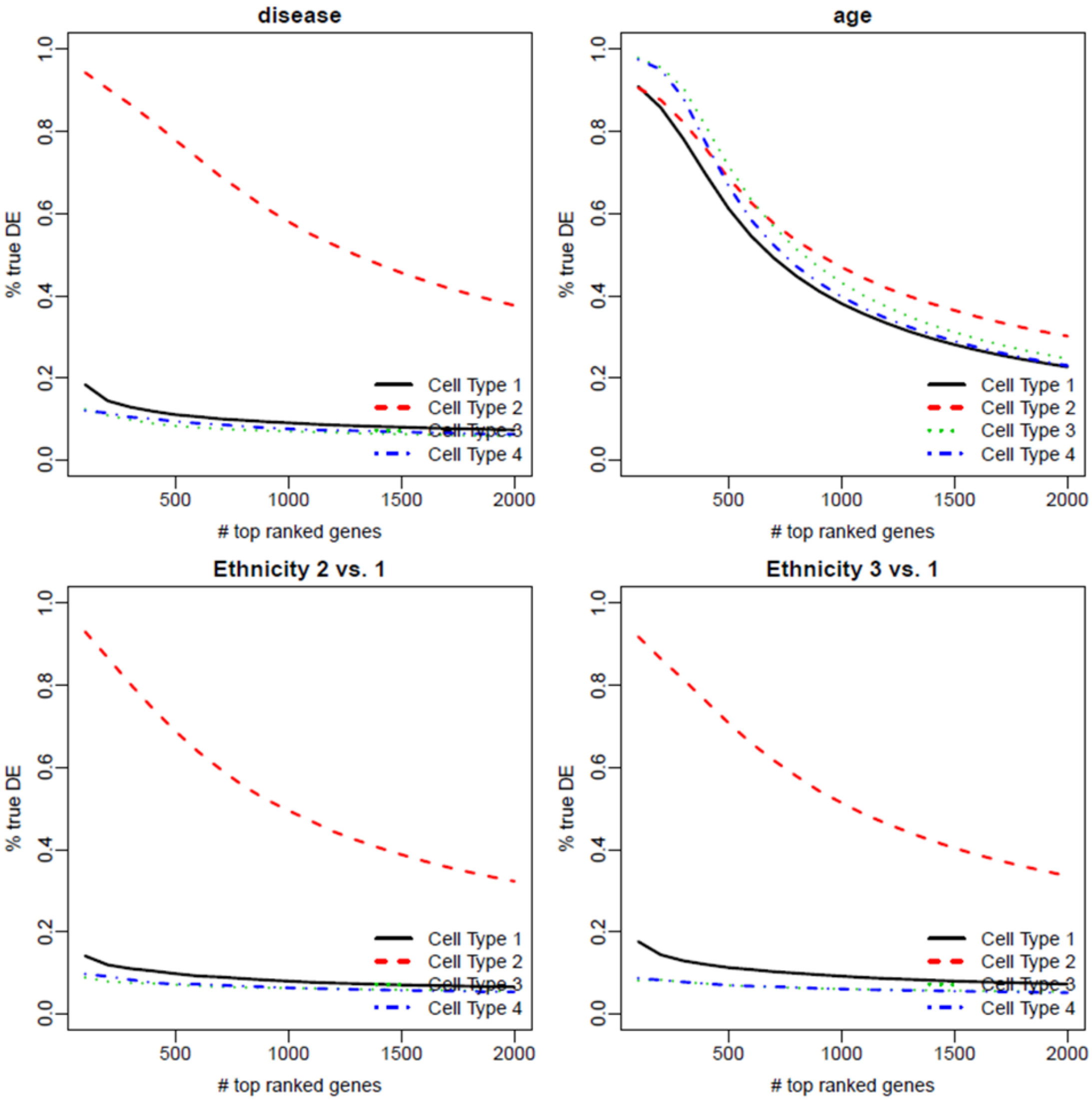


Figure 3c Impact of noise level on DE detection accuracy – n_sd = 10

Figure 2a-2c show the TDR curves for each cell type from TOAST under different sample size. The total sample sized range from 50 to 500, with medium noise level ($n_{sd} = 1$). It is reasonable that larger sample sizes result in better csDE detection. Among four cell types, cell type 2 has the highest accuracy in all four comparisons. When sample size is 100 or larger, TOAST has very good performance with high accuracy for the first 500 top-ranked genes of Cell Type 2. This is reasonably since cell type 2 has the highest proportion in the mixture (as shown in Figure 1). Such phenoemenan also happens when varying the noise levels (Figure 3).

Figure 3a-3c show the TDR curves for each cell type from TOAST with different noise levels. The noise levels range from low ($n_{sd} = 0.1$) to high ($n_{sd} = 10$). We chose a medium sample size (N = 100) in these simulations. We can see that for medium or low noise levels, TOAST performs well. When noise level is high ($n_{sd} = 10$, meaning that the noise level is 10 times of that from the real data), the performance of TOAST is extremely affected.

Overall, the simulation results demonstrate that the method we implemented in TOAST can effectively detect csDE for different type of covariates including continuous and categorical ones.

Real Data

Descriptions of samples in the datasets

Table 1a: Descriptive Statistics of GSE42861

Variables	
Age (years)	55(20-70)
Gender	
· Female	239(71.3%)
· Male	96(28.7%)
Smoking status*	
· Never	101(30.3%)
· Former smoker	108(32.4%)
· Occasional	35(10.55%)
· Current Smoker	89(26.7%)

Total sample size N = 689

* Missing data n = 2

Table 1b: Descriptive Statistics of GSE40279

Variables	
Age (years)	65(19-101)
Gender	
· Female	338(51.5%)
· Male	318(48.5%)
Ethnicity	
· Caucasian - European	426(64.9%)
· Hispanic - Mexican	230(35.1%)

Total sample size N = 656

Values are median (range), or n (%).

We first provided some summary statistics for the two datasets used in this study (Table 1). For the first dataset GSE42681, the mean age of subjects is 52.8 (SD = 11.5), and the range of age is 20 to 70; 239 (71.3%) subjects are female; four types smoking status are reported as Never, Former Smoker, Occasional Smoker, Current Smoker. For the second dataset, the mean of age is 64.0 (SD = 14.7); 338 (51.5%) subjects are female; 426 (64.9%) subjects are Caucasian – European.

Reference-based deconvolution results

EpiDISH with robust partial correlations (RPC) (Teschendorff et al. 2017) was applied to deconvolute DNAm data and to obtain estimated proportions of six cell types based on 10000 marker CpGs selected in each cell type. Figure 4 shows the estimated proportions of each cell type in normal subjects from two datasets. The estimated proportions are similar between the two datasets, and granulocytes occupies the highest proportion.

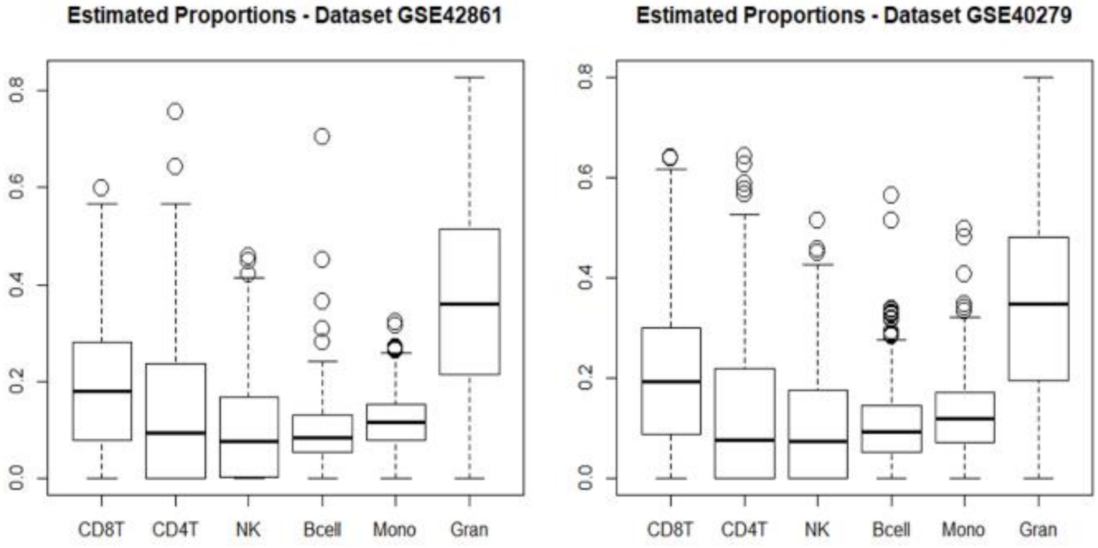


Figure 4 Boxplot of the estimated proportions in real data

Age Effect on Cell Type Proportions

We first examined whether the cell mixing proportions are correlated with age. Estimated cells proportions were plotted against age for each cell type in Figure 5 and Figure 6. The blue lines are fitted LOESS curves and confidence intervals in grey regions. To test the correlation between cell proportions and subject age, we applied the simple linear regression to test the coefficients of age and reported Spearman correlation coefficients for each cell type. From most tests, we found weak but statistically significant association between age and cell type proportions. In GSE42681 dataset, with age increasing, the proportions of CD8T (p-value = 0.0002), NK (p-value = 0.0015) and monocytes (p-value = 0.0012) increase, while the proportions of CD4T (p-value = 0.0001) and granulocytes (p-value = 0.0106) decrease. In GSE40279 dataset, the proportions of CD8T (p-value < 0.0001), monocytes (p-value < 0.0001) and granulocytes (p-value <-0.0001) increase while the proportions of CD4T (p-value < 0.0001), NK (p-value < 0.0001) and B cells (p-value < 0.0001) decrease. These results basically agree with previous findings (Jaffe and Irizarry 2014).

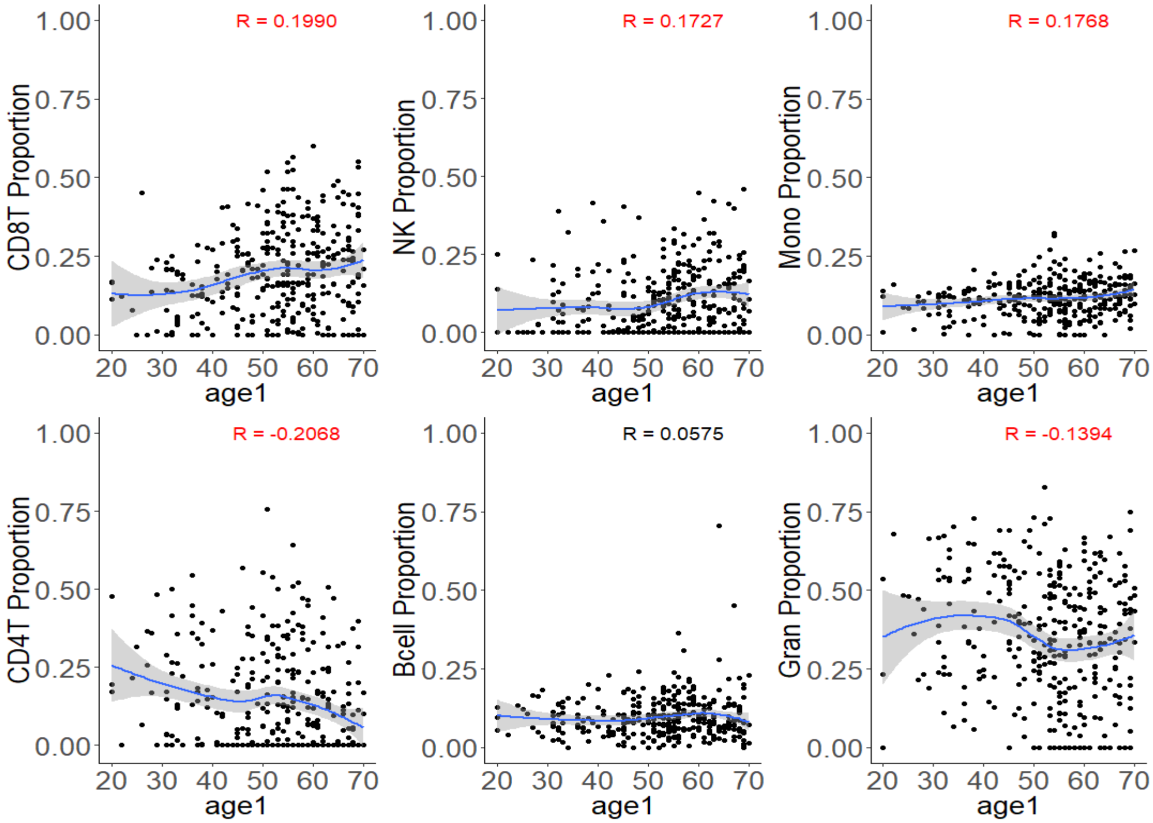


Figure 5 Cell proportions changes across the age - GSE42681

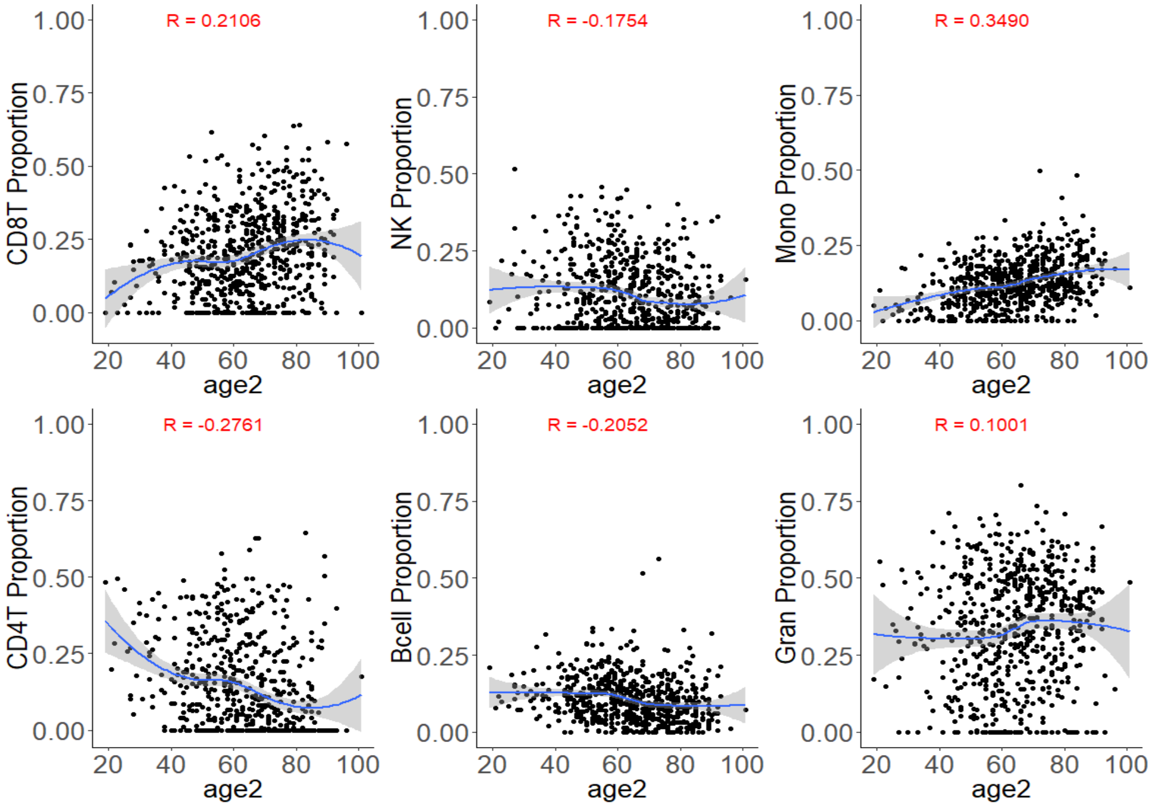


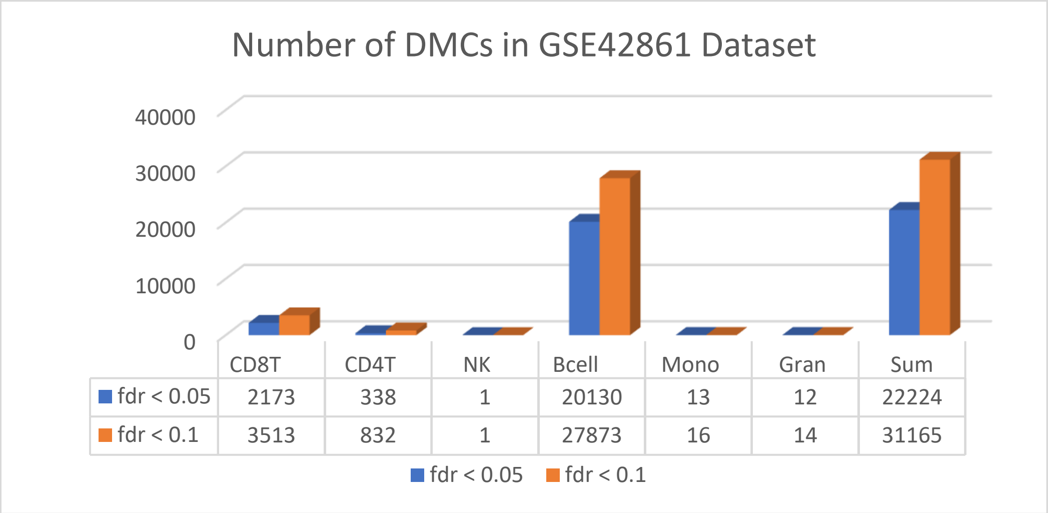
Figure 6 Cell proportions changes across the age - GSE40279

Age Effect on CpGs

We used TOAST to detect cell-type specific changes with age. Differential methylated CpG (DMC) were defined as the ones with q-value (FDR) less than 0.05 from the csDM test. For GSE42861 dataset, Figure 7 suggests the numbers of DMCs showing significant association with age vary among different cell types. Most DMCs were found from B cells.

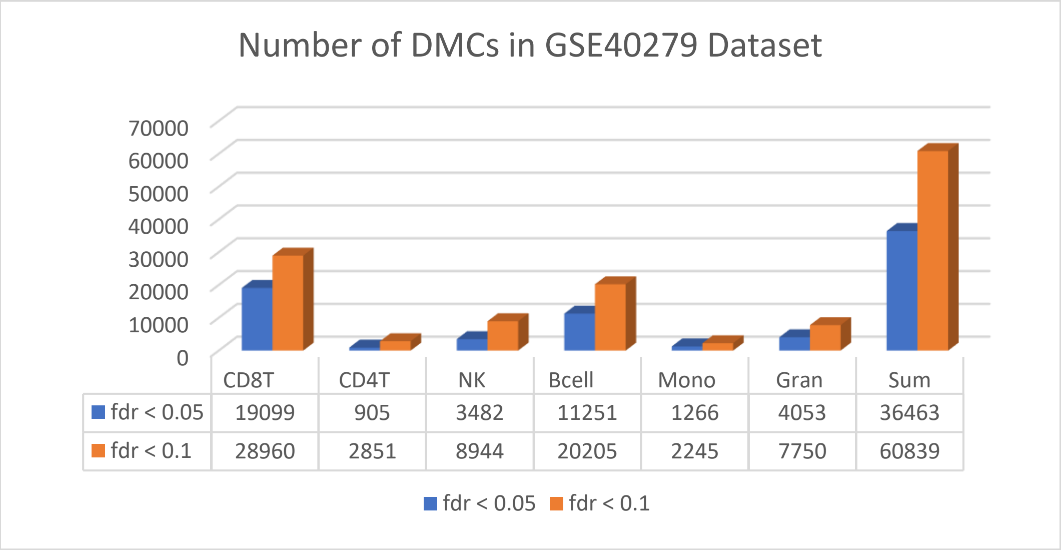
For GSE40279 dataset, the age effect on DMCs also suggests various patterns among cell types. Compared to GSE42681 dataset, more DMCs were found from CD8T cells in GSE40279 （Figure 8）.

Overall, the real data application study suggests that TOAST can detect age-related cell-specific DMCs. To further validate these findings, we compared the csDMCs found from two datasets. Table 2 shows the numbers of cell-specific and total overlapped CpGs (with FDR < 0.05) in two datasets and Chi-square test p-values for testing the overlaps. There results show that the number of overlapped CpGs which associate with age are significant. Most DMCs exist in the CD8T and B cells (Figure 9), which suggests the results of two datasets are in line with each other.



	CD8T	CD4T	NK	Bcell	Mono	Gran	Sum
fdr < 0.05	2173	338	1	20130	13	12	22224
fdr < 0.1	3513	832	1	27873	16	14	31165

Figure 7 Histogram of the number of cell-type specific DMCs - GSE42681



	CD8T	CD4T	NK	Bcell	Mono	Gran	Sum
fdr < 0.05	19099	905	3482	11251	1266	4053	36463
fdr < 0.1	28960	2851	8944	20205	2245	7750	60839

Figure 8 Histogram of the number of cell-type specific DMCs - GSE40279

Table 2: Number of Overlapped DMCs between Two Datasets

	CD8T	CD4T	NK	Bcell	Mono	Gran	Sum
Number of overlapped DMCs	1613	37	0	2652	0	5	4264
P-value	< 0.001	< 0.001	-	< 0.001	-	< 0.001	< 0.001

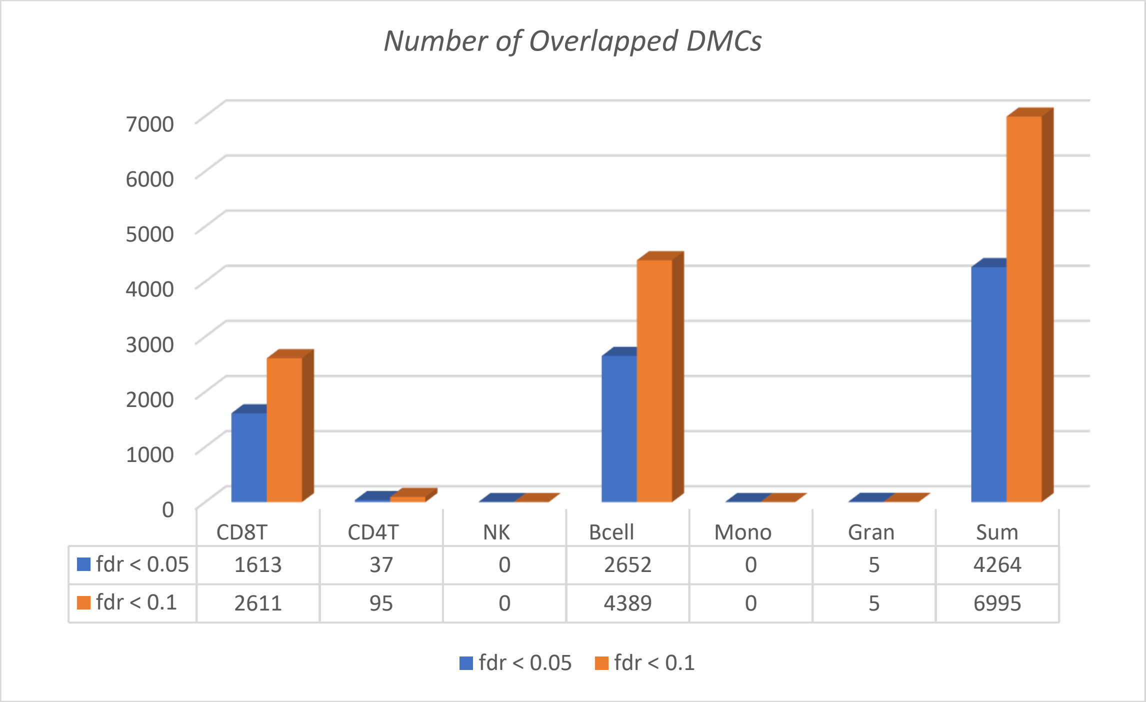


Figure 9 Histogram of the number of overlapped cell-type specific DMCs between GSE42681 and GSE40279

As a comparison, we also performed analysis without considering the cell mixture. For that we ran a simple linear regression for each CpG, using age as predictor and beta value as outcome. In GSE42681 dataset, 13,280 DMCs are found; and in GSE40279, 40,915 DMCs are found from this analysis. There are 6502 overlapped DMCs in both two datasets (Chi-square test p-value < 0.001). Furthermore, we fitted a linear regression model with both age and cells proportions as predictors, but only tested the age effect. This is a typical procedure used before for adjusting the cell mixture in EWAS. From this, fewer DMCs are found (3,912 in GSE42681 and 21,095 in GSE40279), there are 3,556 DMCs overlapped in two datasets (Chi-square test p-value < 0.001). In addition, there are 567 overlapped DMCs datasets (Chi-square test p-value < 0.001) in both TOAST and linear regression model in GSE42681 and 8825 overlapped DMCs datasets (Chi-square test p-value < 0.001) in GSE40279. Both age and cells proportions affect DNAm. Compared the total number of csDMCs found by TOAST, fewer DMCs are detected in mixed samples with linear regression. This disparity suggests that there are cell type-specific changes of the age impact on CpGs methylation, under different patterns among cell types. TOAST can detect these csDMCs effectively.

Table 3: Number of DMCs Found through TOAST and Linear Regression Model

	GSE42681	**GSE40279**	**Overlaps between Two Datasets**
Total DMCs found through TOAST	22224	36463	6632
DMCs found through linear regression model fixed age and proportions	3912	21095	3556
Overlaps between TOAST and Linear Regression Model	567	8825	

Pathway Analysis based on DMCs

In enrichment analysis, the 4264 DMCs found overlapped in both two datasets are mapped to their nearest genes. GO analysis is conducted by the Enrichr, and Table 2 shows the result of top 20 GO progresses with the highest combined score. More detailed understanding for the biological implication of these results, as well as the enrichment analysis for csDMC is the work we plan to further investigate in the near future.

Table 4: Enriched GO Terms for Overlapped DMCs

GOID	Term	Adjusted P-value	Combined Score
GO:0007399	Nervous system development	0.0000	35.62
GO:2000177	Regulation of neural precursor cell proliferation	0.0016	28.74
GO:0035270	Endocrine system development	0.0016	25.95
GO:0045944	Positive regulation of transcription from RNA polymerase II promoter	0.0021	22.07
GO:0006357	Regulation of transcription from RNA polymerase II promoter	0.0002	21.82
GO:0030182	Neuron differentiation	0.0020	21.07
GO:0045595	Regulation of cell differentiation	0.0071	20.21
GO:0001938	Positive regulation of endothelial cell proliferation	0.0011	20.14
GO:0048562	Embryonic organ morphogenesis	0.0016	19.28
GO:0045110	Intermediate filament bundle assembly	0.1532	19.25
GO:0032000	Positive regulation of fatty acid beta-oxidation	0.1725	18.96
GO:0045893	Positive regulation of transcription, DNA-templated	0.0037	18.67
GO:2000544	Regulation of endothelial cell chemotaxis to fibroblast growth factor	0.1725	18.35
GO:0032526	Response to retinoic acid	0.0167	18.27
GO:0048663	Neuron fate commitment	0.0434	17.40
GO:0071300	Cellular response to retinoic acid	0.0269	17.15
GO:0048699	Generation of neurons	0.0021	16.79
GO:0045073	Regulation of chemokine biosynthetic process	0.1725	16.55
GO:0060429	Epithelium development	0.0261	16.51

Discussion

In this work, we extended a method (TOAST) to model the high-throughput data from mixed, heterogeneous samples. TOAST is able to conduct deconvolution on mixed signals and performs hypothesis testing on cell-specific or joint signals. In previous study, TOAST was used to detect csDE/csDM under two-group comparison (Li et al. 2018). In this work, we upgraded TOAST to work for multiple level categorical covariates and continuous covariate age.

We performed extensive simulation studies to evaluate the implementation. We found the sample size has positive effect on performance accuracy, while the noise level has negative effect. Furthermore, our results show that cell type with higher proportions will have better csDE results on almost all occasions. Among three covariates (disease, ethnicity, and age), the TDR curves for each cell type are closer together in age panel than other two because of the marker gene selection and the regulation levels. For categorical covariates disease and ethnicity, we randomly chose one set of marker genes for each group in each covariate, for instance, ethnicity 1 and ethnicity 2 have different marker genes. For continuous covariate age, we only selected one set of marker genes for regulation. Moreover, we drew the up- or down-regulation for disease and ethnicity from N(±1, 0.2). We applied N(±0.03, 0.1) for per unit age, and the regulation scales are proportional to the age. Thus, compared with baseline subjects, the variation scale for age is different from other two covariates.

We applied the upgraded TOAST on two real DNA methylation datasets to look for cell type specific CpG sites associated with age. We first found that cell type proportions are associated with age. For instance, the proportion of CD4+ cells continues to decline with aging, which associated with the loss of thymic tissue (Lynch et al. 2009). Through csDM analysis using TOAST, we identified a number of age-related DMCs in each cell type. While the previous study reported that DNA methylation associates with the long-term scaled age (Teschendorff, West, and Beck 2013), we found that the numbers of age-related csDMCs are different in each of the six cell types. Since the DNAm profiles differ among blood cell types and cell proportions change with aging, cell proportion is reported as an unignorable source of variability in DNAm (Jaffe and Irizarry 2014). In our analysis without considering the cell mixture, we applied two simple linear regression models for each CpGs, one fitted the age effect and another one fitted both age and cell proportions and detected the age-related DMCs of two datasets. There are 6,502 overlapped DMCs found in both two datasets through the model only fitted age, while there are only about half number of overlapped DMCs detected through the model fitted both age and proportions. We could conjecture that cell proportions introduce variations into the global DNAm profiles. Overall, both age and cell proportions could be effect modifiers of DNAm, which worth more biological and pathological investigation.

The aim of TOAST is to detect cell-specific DE/DM in complex tissues. This study focuses on high-throughput experiments data, gene expression data in simulation study and DNA methylation data in real data application. In the future, TOAST could be extended for other types of high-throughput data such as proteomics and metabolomics.

Reference

Absher, Devin M, Xinrui Li, Lindsay L Waite, Andrew Gibson, Kevin Roberts, Jeffrey Edberg, W Winn Chatham, and Robert P Kimberly. 2013. "Genome-Wide DNA Methylation Analysis of Systemic Lupus Erythematosus Reveals Persistent Hypomethylation of Interferon Genes and Compositional Changes to CD4+ T-Cell Populations." *PLOS Genetics* 9 (8): e1003678.

Achim, Kaia, Jean-Baptiste Pettit, Luis R Saraiva, Daria Gavriouchkina, Tomas Larsson, Detlev Arendt, and John C Marioni. 2015. "High-Throughput Spatial Mapping of Single-Cell RNA-Seq Data to Tissue of Origin." *Nature Biotechnology* 33 (5): 503.

Bell, Jordana T, Pei-Chien Tsai, Tsun-Po Yang, Ruth Pidsley, James Nisbet, Daniel Glass, Massimo Mangino, et al. 2012. "Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population." *PLoS Genetics* 8 (4): e1002629–e1002629.

Bird, Adrian. 2002. "DNA Methylation Patterns and Epigenetic Memory." *Genes & Development* 16 (1): 6–21.

Brady, Gerard, Filio Billia, Jennifer Knox, Trang Hoang, Ilan R. Kirsch, Evelyn B. Voura, Robert G. Hawley, et al. 1995. "Analysis of Gene Expression in a Complex Differentiation Hierarchy by Global Amplification of CDNA from Single Cells." *Current Biology* 5 (8): 909–22.

Chen, Edward Y., Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela V. Meirelles, Neil R. Clark, and Avi Ma'ayan. 2013. "Enrichr: Interactive and

Collaborative HTML5 Gene List Enrichment Analysis Tool.” *BMC Bioinformatics* 14.

Christensen, Brock C., Joseph L. Wiemels, Heather H. Nelson, Ru-Fang Yeh, Shichun Zheng, Margaret R. Wrensch, Karl T. Kelsey, et al. 2009. “Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context.” *PLoS Genetics* 5 (8): e1000602.

Clarke, Robert, Habtom W Ressom, Antai Wang, Jianhua Xuan, Minetta C Liu, Edmund A Gehan, and Yue Wang. 2008. “The Properties of High-Dimensional Data Spaces: Implications for Exploring Gene and Protein Expression Data.” *Nature Reviews Cancer* 8 (January): 37.

Dalerba, Piero, Tomer Kalisky, Debashis Sahoo, Pradeep S Rajendran, Michael E Rothenberg, Anne A Leyrat, Sopheak Sim, et al. 2011. “Single-Cell Dissection of Transcriptional Heterogeneity in Human Colon Tumors.” *Nature Biotechnology* 29 (November): 1120.

Erkkilä, Timo, Saara Lehmusvaara, Pekka Ruusuvuori, Tapio Visakorpi, Ilya Shmulevich, and Harri Lähdesmäki. 2010. “Probabilistic Analysis of Gene Expression Measurements from Heterogeneous Tissues.” *Bioinformatics* 26 (20): 2571–77.

Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, et al. 2013. “Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates.” *Molecular Cell* 49 (2): 359–67.

Heintzman, Nathaniel D, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander

Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, and Christina W Ching. 2009. “Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression.” *Nature* 459 (7243): 108.

Ho, Samuel B, Gloria A Niehans, Carolyn Lyftogt, Pei Sha Yan, David L Cherwitz, Elizabeth T Gum, Rajvir Dahiya, and Young S Kim. 1993. “Heterogeneity of Mucin Gene Expression in Normal and Neoplastic Tissues.” *Cancer Research* 53 (3): 641 LP-651.

Horvath, Steve. 2016. “DNA Methylation Age of Human Tissues and Cell Types,” 1–4.

Houseman, Eugene A., William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. 2012. “DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution.” *BMC Bioinformatics* 13 (1).

Hu, Ping, Wenhua Zhang, Hongbo Xin, and Glenn Deng. 2016. “Single Cell Isolation and Analysis.” *Frontiers in Cell and Developmental Biology* 4 (October): 1–12.

Jabbari, Kamel, and Giorgio Bernardi. 2004. “Cytosine Methylation and CpG, TpG (CpA) and TpA Frequencies.” *Gene* 333 (SUPPL.): 143–49.

Jaffe, Andrew E, and Rafael A Irizarry. 2014. “Accounting for Cellular Heterogeneity Is Critical in Epigenome-Wide Association Studies.” *Genome Biology* 15 (2): R31–R31.

Jin, Bilian, Yajun Li, and Keith D Robertson. 2011. “DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?” *Genes & Cancer* 2 (6): 607–17.

Kuhn, Alexandre, Azad Kumar, Alexandra Beilina, Allissa Dillman, Mark R. Cookson, and Andrew B. Singleton. 2012. “Cell Population-Specific Expression Analysis of Human Cerebellum.” *BMC Genomics* 13 (1).

Kuhn, Alexandre, Doris Thu, Henry J Waldvogel, Richard L M Faull, and Ruth Luthi-Carter. 2011. “Population-Specific Expression Analysis (PSEA) Reveals Molecular Changes in Diseased Brain.” *Nature Methods* 8 (October): 945.

Kuleshov, Maxim V, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. “Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update.” *Nucleic Acids Research* 44 (W1): W90–97. https://doi.org/10.1093/nar/gkw377.

Levine, Morgan E., H. Dean Hosgood, Brian Chen, Devin Absher, Themistocles Assimes, and Steve Horvath. 2015. “DNA Methylation Age of Blood Predicts Future Onset of Lung Cancer in the Women’s Health Initiative.” *Aging* 7 (9): 690–700.

Li, Ziyi, Zhijin Wu, Jin Peng, and Hao Wu. 2018. “Dissecting Differential Signals in High-Throughput Data from Complex Tissues.” *BioRxiv Bioinformatics*, 1–8.

Liu, Yun, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, et al. 2013. “Epigenome-Wide Association Data Implicate DNA Methylation as an Intermediary of Genetic Risk in Rheumatoid Arthritis” 31 (2): 142–47.

Lynch, Heather E, Gabrielle L Goldberg, Ann Chidgey, Marcel R M Van den Brink, Richard Boyd, and Gregory D Sempowski. 2009. “Thymic Involution and Immune Reconstitution.” *Trends in Immunology* 30 (7): 366–73.

Mäkinen, Veli, Djamal Belazzougui, Fabio Cunial, and Alexandru I Tomescu. 2015. *Genome-Scale Algorithm Design*. Cambridge University Press.

Newman, Aaron M, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. 2015. "Robust Enumeration of Cell Subsets from Tissue Expression Profiles." *Nature Methods* 12 (5): 453.

Rakyan, Vardhman K, Thomas A Down, Siarhei Maslau, Toby Andrew, Tsun-Po Yang, Huriya Beyan, Pamela Whittaker, Owen T McCann, Sarah Finer, and Ana M Valdes. 2010. "Human Aging-Associated DNA Hypermethylation Occurs Preferentially at Bivalent Chromatin Domains." *Genome Research* 20 (4): 434–39.

Shen-Orr, Shai S., and Renaud Gaujoux. 2013. "Computational Deconvolution: Extracting Cell Type-Specific Information from Heterogeneous Samples." *Current Opinion in Immunology* 25 (5): 571–78.

Shen-Orr, Shai S., Robert Tibshirani, Purvesh Khatri, Dale L. Bodian, Frank Staedtler, Nicholas M. Perry, Trevor Hastie, Minnie M. Sarwal, Mark M. Davis, and Atul J. Butte. 2010. "Cell Type-Specific Gene Expression Differences in Complex Tissues." *Nature Methods* 7 (4): 287–89.

Shipitsin, Michail, Lauren L. Campbell, Pedram Argani, Stanislawa Weremowicz, Noga Bloushtain-Qimron, Jun Yao, Tatiana Nikolskaya, et al. 2007. "Molecular Definition of Breast Tumor Heterogeneity." *Cancer Cell* 11 (3): 259–73.

Sonnen, Joshua A, Eric B Larson, Sebastien Haneuse, Randy Woltjer, Ge Li, Paul K Crane, Suzanne Craft, and Thomas J Montine. 2009. "Neuropathology in the Adult

Changes in Thought Study: A Review." *Journal of Alzheimer's Disease* 18 (3): 703–11.

Teschendorff, Andrew E., Charles E. Breeze, Shijie C. Zheng, and Stephan Beck. 2017. "A Comparison of Reference-Based Algorithms for Correcting Cell-Type Heterogeneity in Epigenome-Wide Association Studies." *BMC Bioinformatics* 18 (1): 1–14.

Teschendorff, Andrew E, James West, and Stephan Beck. 2013. "Age-Associated Epigenetic Drift: Implications, and a Case of Epigenetic Thrift?" *Human Molecular Genetics* 22 (R1): R7–15.

Venet, D., F. Pecasse, C. Maenhaut, and H. Bersini. 2001. "Separation of Samples into Their Constituents Using Gene Expression Data." *Bioinformatics* 17 (SUPPL. 1): 279–87.

Walker, Michael D, Thomas Edlund, Anne M Boulet, and William J Rutter. 1983. "Cell-Specific Expression Controlled by the 5′-Flanking Region of Insulin and Chymotrypsin Genes." *Nature* 306 (5943): 557.

Wang, Sun Chong, Beatrice Oeize, and Axel Schumacher. 2008. "Age-Specific Epigenetic Drift in Late-Onset Alzheimer's Disease." *PLoS ONE* 3 (7).

Whitney, A. R., M. Diehn, S. J. Popper, A. A. Alizadeh, J. C. Boldrick, D. A. Relman, and P. O. Brown. 2003. "Individuality and Variation in Gene Expression Patterns in Human Blood." *Proceedings of the National Academy of Sciences* 100 (4): 1896–1901.

Wolslegel, Kristen, Dhaya Seshasayee, Zora Modrusan, Hilary F. Clark, and Alexander

R. Abbas. 2009. “Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus.” *PLoS ONE* 4 (7): e6098.

Zheng, Shijie C, Amy P Webster, Danyue Dong, Andy Feber, David G Graham, Roisin Sullivan, Sarah Jevons, Laurence B Lovat, Stephan Beck, and Martin Widschwendter. 2018. “A Novel Cell-Type Deconvolution Algorithm Reveals Substantial Contamination by Immune Cells in Saliva, Buccal and Cervix.” *Epigenomics* 10 (7): 925–40.