Dissecting differential signals in high-throughput data from complex tissues

Supplementary materials

Ziyi Li[1], Zhijin Wu[2], Peng Jin[3] and Hao Wu[1,*]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA.

[2]Department of Biostatistics, Brown University, Providence, RI 02806, USA

[3]Department of Human Genetics, Emory University, Atlanta, GA 30322, USA

* Correspondance: hao.wu@emory.edu.

March 27, 2019

S1. Procedures of calculating lfc, absolute diff, and csSAM results

Given the observations $\boldsymbol{Y}$ and (estimated) mixture proportions $\boldsymbol{\theta}$, we first estimate the pure tissue profiles for cases $(\hat{\boldsymbol{x}}_{\boldsymbol{1}}, \cdots, \hat{\boldsymbol{x}}_{\boldsymbol{K}})$ and controls $(\hat{\boldsymbol{x}}^{*}_{\boldsymbol{1}}, \cdots, \hat{\boldsymbol{x}}^{*}_{\boldsymbol{K}})$ using the `basis` function from Bioconductor package *csSAM*. For cell type k,

1. the lfc (log fold change) of cases and controls are computed by $lfc_k = abs(log(\hat{\boldsymbol{x}}_{\boldsymbol{k}}) - log(\hat{\boldsymbol{x}}^{*}_{\boldsymbol{k}}))$, $k = 1, \cdots, K$. $abs()$ is the absolute operation.

2. the absolute diff (absolute difference) of cases and controls are computed by $absolute\ diff_k = abs(\hat{\boldsymbol{x}}_{\boldsymbol{k}} - \hat{\boldsymbol{x}}^{*}_{\boldsymbol{k}})$, $k = 1, \cdots, K$.

3. the csSAM results are obtained using the `csSamWrapper` function from the *csSAM* package with $\boldsymbol{Y}$, $\boldsymbol{\theta}$ and sample size as inputs.

S2. Procedures of testing for other hypothesis

Most simulation results prsented in the main manuscript are designed to test Hypothesis 1 in Section 2.2. To demonstrate the functionalities to test other hypothesis, we conduct additional simulations using the following procedures.

We use the same simulation model as described in Section 2.3. However, different from testing one cell type between cases and controls, DE signals already exist cross different cell types since all pure tissue profiles are generated based on real dataset. Instead, we apply Bioconductor package *limma* on underlying true pure tissue profiles to define which genes are true DE's. Following the notation in Section 2.3, the detailed procedures are as follows:

- Testing for Hypothesis 2 in Section 2.2

    1. For subject i from the control group, denote the individualized reference panel as X_i, which is a G by K matrix.
    2. For cell type p and q $(p, q = 1, \cdots, K, p \neq q)$, extract all the $p-$th and $q-$th columns from $X_i, i = 1, \cdots, s_1$ to a new matrix $X^{(p,q)}$. Apply *limma* on $X^{(p,q)}$ and define all genes with $FDR < 0.05$ and absolute log fold change > 3 as true DE genes between cell types p and q in controls.
    3. Apply the proposed method on observed data $Y_{control}$ with all the columns corresponding to control subjects and obtain the lists of adjusted p values for all cell types.
    4. Use true DE status from step 2 and p values from step 3 to draw TDR curves as in Figure S7(a)

- Testing for Hypothesis 3 in Section 2.2

    1. For subject j from the disease group, denote the individualized reference panel as X^*_j, which is a G by K matrix.
    2. For cell type p and q $(p, q = 1, \cdots, K, p \neq q)$, extract all the $p-$th and $q-$th columns from $X^*_j, j = 1, \cdots, s_2$ to a new matrix $X^{*(p,q)}$. Apply *limma* on $X^{*(p,q)}$ and define all genes with $FDR < 0.05$ and absolute log fold change > 3 as true DE genes between cell types p and q in disease group.
    3. Apply the proposed method on observed data $Y_{disease}$ with all the columns corresponding to disease subjects and obtain the lists of adjusted p values for all cell types.
    4. Use true DE status from step 2 and p values from step 3 to draw TDR curves as in Figure S7(b)

- Testing for Hypothesis 4 in Section 2.2

    1. For subject i from the control group, denote the individualized reference panel as X_i, and similarly for subject j from the disease group, denote the individualized reference panel as X^*_j. Both X_i and X^*_j are a G by K matrix.
    2. For cell type p and q $(p, q = 1, \cdots, K, p \neq q)$, extract all the $p-$th and $q-$th columns from $X_i, i = 1, \cdots, s_1$ and $X^*_j, j = 1, \cdots, s_2$ to a new matrix $X^{(p,q)}$. Apply *limma* on $X^{*(p,q)}$ with a design matrix including interaction terms. Define all genes with $FDR < 0.05$ as true DE genes between cell types p and q across control and disease group.
    3. Apply the proposed method on observed data Y with all subjects and obtain the lists of adjusted p values for all comparisons.
    4. Use true DE status from step 2 and p values from step 3 to draw TDR curves as in Figure S7(c)

S3. Simulation study for RNA-seq data

We conducted simulation studies for detecting cell type specific differeneital expression for RNA-seq data. The simulation parameters are estimated from a real dataset on GEO with accession number GSE60424 [1]. This datasets has RNA-seq measurements of 6 purified blood cells, with 4 replicates for controls and for patients with an array of immune-associated diseases (type I diabetes, amyotrophic lateral sclerosis, sepsis, and multiple sclerosis patients).

The detail simulation steps are listed below:

1. For cell type k and gene g, estimate mean μ_g^k and dispersion ϕ_g^k from the control group using the `estParam` function in Bioconductor package *PROPER* [2].

2. For cases, we assume they have the same gene-specific dispersons ϕ_g^k but different means for csDE genes. The mean expressionss for cases, denoted by $\mu_g^{k'}$, is generated based on μ_g^k with log fold changes (lfc) added to a randomly selected group of DE genes. Lfc is simulated from $0.5N(2,1) + 0.5N(2,1)$.

3. For person i, simulate the individualized underlying cell type specific expressions m_{gi}^k from a Gamma distribution.

$$m_{gi}^k \sim \Gamma(\mu_g^k, \mathbf{scale} = \frac{\phi_g^k}{1-\phi_g^k})\ for\ controls$$

and

$$m_{gi}^k \sim \Gamma(\mu_g^{k'}, \mathbf{scale} = \frac{\phi_g^k}{1-\phi_g^k})\ for\ cases$$

4. For person i, mix the cell type specific expressions to generate the expression for mixed samples. The mixing proportions are generated using the same procedures presented in the main manuscript.

$$m_{gi} \sim \sum_k \pi_i^k m_{gi}^k$$

5. Generate read count Y_{gi} from Poisson distribution.

$$Y_{gi}|m_{gi} \sim Poisson(m_{gi})$$

After observations $Y = (Y_{gi})$ are generated, we apply the proposed method on the raw counts, and the results are presented in Figure S8. The results are summarized over 20 Monte Carlo datasets.

S4. Fitting the model in original or log scale

Historically, gene expression microarray data analyses are primarily performed in logarithm scale. Under the sample mixing context, the mixing of pure cell type expression values is believed to take place in the raw scale, and the linear relationship could be destroyed if one log-transforms the data. This is also the reason why many proportion estimation methods (either reference-free or reference-based) are performed on the raw instead of log-transformed data [3, 4, 5, 6]. However there is also voice for using log-scale data [7].

In our simulation study, we evaluate the impact of log-transformation on csDE detection. Figure S5 compares the TDR curves from TOAST for using raw vs. log transformed data in csDE detection for all cell types, using both reference-based and reference free proportion estimates. The result is a mixed performance without a universal benefit for using either the log or original scale. Comparison using the Immune dataset (Figure S11) also shows similar performance on both scales.

In simulations we try to mimic data from real biological experiments as much as possible. However, real data sets often contain more complex noise sources and outliers not captured in simulations. Log transformation provides a natural regularization against such noise and outliers, which may lead to more robust analysis in real data applications. We hold a neural position given the current evidence except to remind users that the interpretation of the differential effect sizes should depend on the analysis scale: the estimated difference is a relative value (log fold change) in log scale, whereas in the raw scale it may need to be interpreted in the context of a feature's baseline expression/methylation level.

S5. Application to HIV DNA methylation data

We have also applied the proposed method to a DNA methylation dataset downloaded from GEO with accession number GSE67705 [8]. Unlike the brain dataset, this HIV dataset only contains measurements for whole blood, thus we benchmark the results using enrichment analysis and literature research. We obtain whole blood 450K DNA methylation measurements from 142 HIV patients and 44 healthy controls. The DNA methylation blood reference is obtained from Bioconductor package *FlowSorted.Blood.450k* [9]. The batch effect is removed by Bioconductor package *sva*. We solve the mixture proportions by *EpiDISH* from Teschendorff et al BMC Bioinformatics 2017 [10], which is a reference-based deconvolution software designed for DNA methylation data. After applying TOAST we find the following signals, we find Monocyte has the largest number of DMC. There are 26 DMCs with $FDR < 0.05$, 731 DMCs with $FDR < 0.2$, and 2987 with $pvalue < 0.01$. These 2987 DMC can be mapped to 755 genes and we conduct anenrichment analysis using software EnrichR (http://amp.pharm.mssm.edu/Enrichr/enrich).

The dbGap (the database of Genotypes and Phenotypes) enrichment analysis results are presented in Figure S12. We find that Lymphocytes and HIV-1 have been identified as top terms, which may suggest the validity of our findings. Besides enrichment analysis, we also find a few recent publications that highlighted the important role of Monocytes in AIDS [11, 12, 13]. Especially the last paper by Hasegawa et al. finds that the level of monocyte turnover was not linked to the CD4+ T-cell count and was a better predictive marker for AIDS progression than was viral load or lymphocyte activation, which emphasize the important role of Monocytes in AIDS etiology and progression.

S6. Several existing methods are special cases of the proposed method

A number of existing methods are available for detecting cell-type specific differential signals. Here we discuss the connections between three existing methods (cell-type specific differential methylation in human brain tissue [14], referred to as csDMHB; population-specific expression analysis (PSEA) [15]; and cell specific eQTL analysis [16], referred to as cseQTL) and the proposed method. We show that they are simplified version or special cases of our framework in testing cell-type specific differential signals. Note that none of these existing methods provide functionality for flexible hypothesis test, such as testing the difference among cell types in one condition as we showed in the Immune Dataset example.

Connection with csDMHB

csDMHB presented a statistical model that estimates differences in DNA methylation between two brain regions dorsolateral prefrontal cortex (D) and Hippocampal formation (H). Assume there are two cell types(NeuN+ and NeuN-) in the brain, they want to detect whether cell-type specific differences exist between the two brain regions. Following the notation from [14], Y_i is the observation for subject i. Indicator variable $X_i = 1$ if sample i from H and $X_i = 0$ otherwise. Their model is $E(Y_i) = \mu_{D,+} + (\mu_{D,-} - \mu_{D,+})\pi_i + (\mu_{H,+} - \mu_{D,+})X_i(1-\pi_i) + (\mu_{H,-} - \mu_{D,-})X_i\pi_i$. To re-format this function into a linear model framework, they use $\beta_0, \beta_1, \beta_2, \beta_3$ to represent $\mu_{D,+}$, $\mu_{D,-} - \mu_{D,+}$, $\mu_{H,+} - \mu_{D,+}$, and $\mu_{H,-} - \mu_{D,-}$ respectively.

Comparing csDMHB with our model, csDMHB is actually a simplified version of our framework with condition $K = 2$. Using our notation, when $K = 2$, we have $E(Y_i) = \theta_{i1}(\mu_1 + \beta_1 Z_i) + \theta_{i2}(\mu_2 + \beta_2 Z_i)$. Here Z_i is the indicator function that equals 1 if sample i from H and 0 otherwise. From this step, we can rearrange

the term and build connections with csDMHB:

$$\begin{aligned} E(Y_i) &= \theta_{i1}(\mu_1 + \beta_1 Z_i) + \theta_{i2}(\mu_2 + \beta_2 Z_i) \\ &= \mu_1 + (\mu_2 - \mu_1)\theta_{i2} + \beta_1 Z_i(1-\theta_{i2}) + \beta_2 Z_i \theta_{i2} \quad (1) \\ &= \mu_{D,+} + (\mu_{D,-} - \mu_{D,+})\pi_i + (\mu_{H,+} - \mu_{D,+})X_i(1-\pi_i) + (\mu_{H,-} - \mu_{D,-})X_i\pi_i \quad (2) \\ &= \beta_0 + \beta_1\pi_i + \beta_2 X_i(1-\pi_i) + \beta_3 X_i \pi_i. \end{aligned}$$

Transition from (1) to (2) implies $\beta_0, \beta_1, \pi_i, \beta_2, X_i, \beta_3$ in the notation of csDMHB corresponds to $\mu_1, \mu_2 - \mu_1, \theta_{i2}, \beta_1, Z_i, \beta_2$ in our notation system. They use the hypotheses $\beta_2 = 0$ and $\beta_3 = 0$ to test whether there is difference in NeuN+ methylation between D and H, and whether there is difference in NeuN− methylation between D and H respectively. This aligns with testing $\beta_1 = 0$ and $\beta_2 = 0$, which we proposes to test differences in cell type 1 and 2. Thus csDMHB is a simplified version of our method, for it only considers a mixture with two components. Moreover, it does not provide functionality to test the changes of different cell types under the same condition, e.g., difference between NeuN+ and NeuN− in D or H.

Connection with PSEA

PSEA presents a model formulation that detects the cell-type (or population) specific differential expressed genes with respect to their relative expression level calculated using cell-type specific marker gene expression level. To apply PSEA, one should first obtain a list of cell-type specific marker genes that express in cell-type p^* only. They assume $y = a + \sum_{p=1}^{P} x_p f_p$ where a is background, x_p is the cell-type specific gene expression in cell type p, and f_p is the mixing proportion.

By using cell-type specific marker gene x_{p^*}, $f_{p^*} = (y_{p^*} - a)/x_p$ or its approximation $f_{p^*} = y_{p^*}/x_p$ is used as surrogate for mixture proportion. Thus the PSEA model becomes $y_i = a + \sum_{p=1}^{P} x_p \frac{y_{p^*} - a}{x_{p^*}} = a[1 - \sum \frac{x_p}{x_{p^*}}] + \sum_{p=1}^{P} x_p \frac{y_{p^*}}{x_{p^*}} \approx a + \sum_{p=1}^{P} \beta_p y_{p^*}$. Here $\beta_p = \frac{x_p}{x_{p^*}}$. For two group comparison, PSEA uses an indicator variable d_i with $d_i = 0$ for samples in group 1 and $d_i = 1$ for samples in group 2. The model is represented as

$$y_i = a + \sum_{p,p^*=1}^{P} \beta_p y_{p^*,i} + \sum_{p,p^*=1}^{P} \beta_p^{'} y_{p^*,i} d_i$$

Comparing PSEA to our method, we find the ideas are very similar and our formulation is the generalized version of PSEA. Using our notation,

$$\begin{aligned} E(Y_i) &= \sum_{k=1}^{K} \theta_{ik}\mu_k + \sum_{k=1}^{K} \theta_{ik}\beta_k Z_i \quad (3) \\ &= a + \sum_{p=1}^{P} x_p \cdot \frac{y_{p^*}}{x_{p^*}} + \sum_{p=1}^{P} (x_p^{'} - x_p)\frac{y_p^*}{x_{p^*}} d_i \\ &= a + \sum_{p=1}^{P} \beta_p y_{p^*,i} + \sum_{p=1}^{P} \beta_p^{'} y_{p^*,i} d_i \quad (4) \end{aligned}$$

Transition from (3) to (4) suggests $\beta_p, \beta_p^{'}, d_i, y_{p^*}$ correspond to μ_k/X_{marker}, β_k/X_{marker}, Z_i, $\theta_{ik} \cdot X_{marker}$ respectively. Our method does not have marker gene expressionss in the formula since we assume the mixing proportions θ_{ik} are known, so we just use X_{marker} to symbolically represent the relative relationship. Note that the background term a is assumed omissible in PSEA and is absorbed into μ_1 in our notation system. PSEA uses $H_0 : \beta_p^{'} = 0$ to test the null hypothesis of no difference between the p-th cell types in the two sample groups, which aligns with testing $H_0 : \beta_k = 0$ for detecting the difference between k-th cell types in our notation system.

Similarly in three-group comparison, PSEA has the model

$$y_i = a + \sum_{p,p^*=1}^{P} \beta_p y_{p^*,i} + \sum_{p,p^*=1}^{P} \beta_p^{'} y_{p^*,i} d_{1,i} + \sum_{p,p^*=1}^{P} \beta_p^{''} y_{p^*,i} d_{2,i}.$$

While in our notation system, we only need to specify covariate $\mathbf{Z_i}$ as a vector with $\mathbf{Z_i} = (0,0)^T$for subjects from group 1, $(1,0)^T$ for subjects from group 2, and $(0,1)^T$ for subject from group 3, then the correspondence relationship between PSEA and our model still holds. Thus, PSEA is similar to our method for detecting the expression changes for a specific cell type among different groups. It does not have functionality to test differential expression among different cell types in the same group, or higher order changes.

Connection with cseQTL

cseQTL expands the typical linear model to detect cell-type specific effects using expression quantitative trait loci (eQTL) datasets generated from whole tissue. Instead of using mixture proportions as the previous mentioned methods, cseQTL creates cell-type specific proxy for cell types of interest through correlation-based marker selection process. They treat the cell-type specific proxy as a covariate and add the main effect for the covariate and interaction term of covariate and genotype. Their model can be written as $Y \approx I + \beta_1 \times G + \beta_2 \times P + \beta_3 \times P : G + e$ where Y is the gene expression, G is the genotype information, P is the cell-type specific proxy, $P : G$ is the interaction term between proxy and the genotype. The parameter of interest for testing is β_3.

Comparing the cseQTL model with the existing methods and our proposed method, we find the cseQTL model is very similar to the formulation of csDMGHB and can be seen as a special case of method with $K = 2$. The fact that cseQTL only uses the proxy for one cell type in the formulation implies that they assume only two cell types (cell type of interest and cell type not of interest) in the analysis of each cell type. We can build a connection with cseQTL model as

$$\begin{aligned} E(Y_i) &= \theta_{i1}(\mu_1 + \beta_1 Z_i) + \theta_{i2}(\mu_2 + \beta_2 Z_i) \\ &= \mu_1 + \beta_1 Z_i + (\mu_2 - \mu_1)\theta_{i2} + (\beta_2 - \beta_1) Z_i \theta_{i2} \quad (5) \\ &\approx I + \beta_1 G + \beta_2 P + \beta_3 P : G \quad (6) \end{aligned}$$

We use '$\approx$' to follow the notation utilized in the cseQTL paper [16] which indicates that cell-type proxy is used and the formulation just gives an approximation for single marker $cis-$eQTL mapping. The transition from (5) to (6) implies that we can build correspondence relationship between $\beta_1, \beta_2, \beta_3$ in cseQTL's notation and and $\beta_1, \mu_2 - \mu_1, \beta_2 - \beta_1$ in our notation.

References

[1] Peter S Linsley, Cate Speake, Elizabeth Whalen, and Damien Chaussabel. Copy number loss of the interferon gene cluster in melanomas is linked to reduced t cell infiltrate and poor patient prognosis. PloS one, 9(10):e109760, 2014.

[2] Hao Wu, Chi Wang, and Zhijin Wu. Proper: comprehensive power evaluation for differential expression using rna-seq. Bioinformatics, 31(2):233–241, 2014.

[3] Alexander R Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PloS one, 4(7):e6098, 2009.

[4] Ting Gong, Nicole Hartmann, Isaac S Kohane, Volker Brinkmann, Frank Staedtler, Martin Letzkus, Sandrine Bongiovanni, and Joseph D Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PloS one, 6(11):e27156, 2011.

[5] Yi Zhong, Ying-Wooi Wan, Kaifang Pang, Lionel ML Chow, and Zhandong Liu. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC bioinformatics, 14(1):89, 2013.

[6] Yi Zhong and Zhandong Liu. Gene expression deconvolution in linear space. Nature methods, 9(1):8, 2012.

[7] Renaud Gaujoux. An introduction to gene expression deconvolution and the cellmix package. 2013.

[8] Andrew M Gross, Philipp A Jaeger, Jason F Kreisberg, Katherine Licon, Kristen L Jepsen, Mahdieh Khosroheidari, Brenda M Morsey, Susan Swindells, Hui Shen, Cherie T Ng, et al. Methylome-wide analysis of chronic hiv infection reveals five-year increase in biological age and epigenetic targeting of hla. Molecular cell, 62(2):157–168, 2016.

[9] Flowsorted.blood.450k: Illumina humanmethylation data on sorted blood cell populations.

[10] Andrew E Teschendorff, Charles E Breeze, Shijie C Zheng, and Stephan Beck. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. BMC bioinformatics, 18(1):105, 2017.

[11] Jennifer H Campbell, Anna C Hearps, Genevieve E Martin, Kenneth C Williams, and Suzanne M Crowe. The importance of monocytes and macrophages in hiv pathogenesis, treatment, and cure. Aids, 28(15):2175–2187, 2014.

[12] Tricia H Burdo, Andrew Lackner, and Kenneth C Williams. Monocyte/macrophages and their role in hiv neuropathogenesis. Immunological reviews, 254(1):102–113, 2013.

[13] Atsuhiko Hasegawa, Huining Liu, Binhua Ling, Juan T Borda, Xavier Alvarez, Chie Sugimoto, Heather Vinet-Oliphant, Woong-Ki Kim, Kenneth C Williams, Ruy M Ribeiro, et al. The level of monocyte turnover predicts disease progression in the macaque model of aids. Blood, 114(14):2917–2925, 2009.

[14] Carolina M Montaño, Rafael A Irizarry, Walter E Kaufmann, Konrad Talbot, Raquel E Gur, Andrew P Feinberg, and Margaret A Taub. Measuring cell-type specific differential methylation in human brain tissue. Genome biology, 14(8):R94, 2013.

[15] Alexandre Kuhn, Doris Thu, Henry J Waldvogel, Richard LM Faull, and Ruth Luthi-Carter. Population-specific expression analysis (psea) reveals molecular changes in diseased brain. Nature methods, 8(11):945, 2011.

[16] Harm-Jan Westra, Danny Arends, Tõnu Esko, Marjolein J Peters, Claudia Schurmann, Katharina Schramm, Johannes Kettunen, Hanieh Yaghootkar, Benjamin P Fairfax, Anand Kumar Andiappan, et al. Cell specific eqtl analysis without sorting cells. PLoS genetics, 11(5):e1005223, 2015.

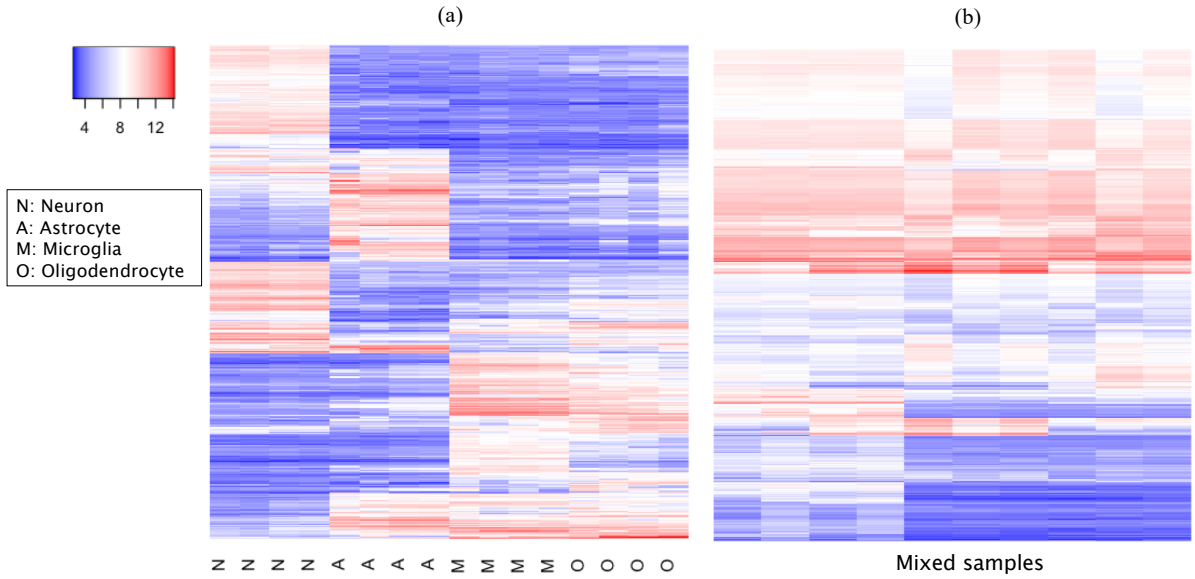


Figure S1: Heatmaps of gene expression profiles for purified rat brain cells and mixed samples from GSE19380. (a) Gene expression from primary brain cell cultures of rat. Purified cell types include neuronal, astrocytic, oligodendrocytic and microglial cultures. (b) Gene expression from RNA mixtures of rat. For both (a) and (b), only the top 1000 most variant genes are presented and the rows have been re-ordered for demonstration purpose.

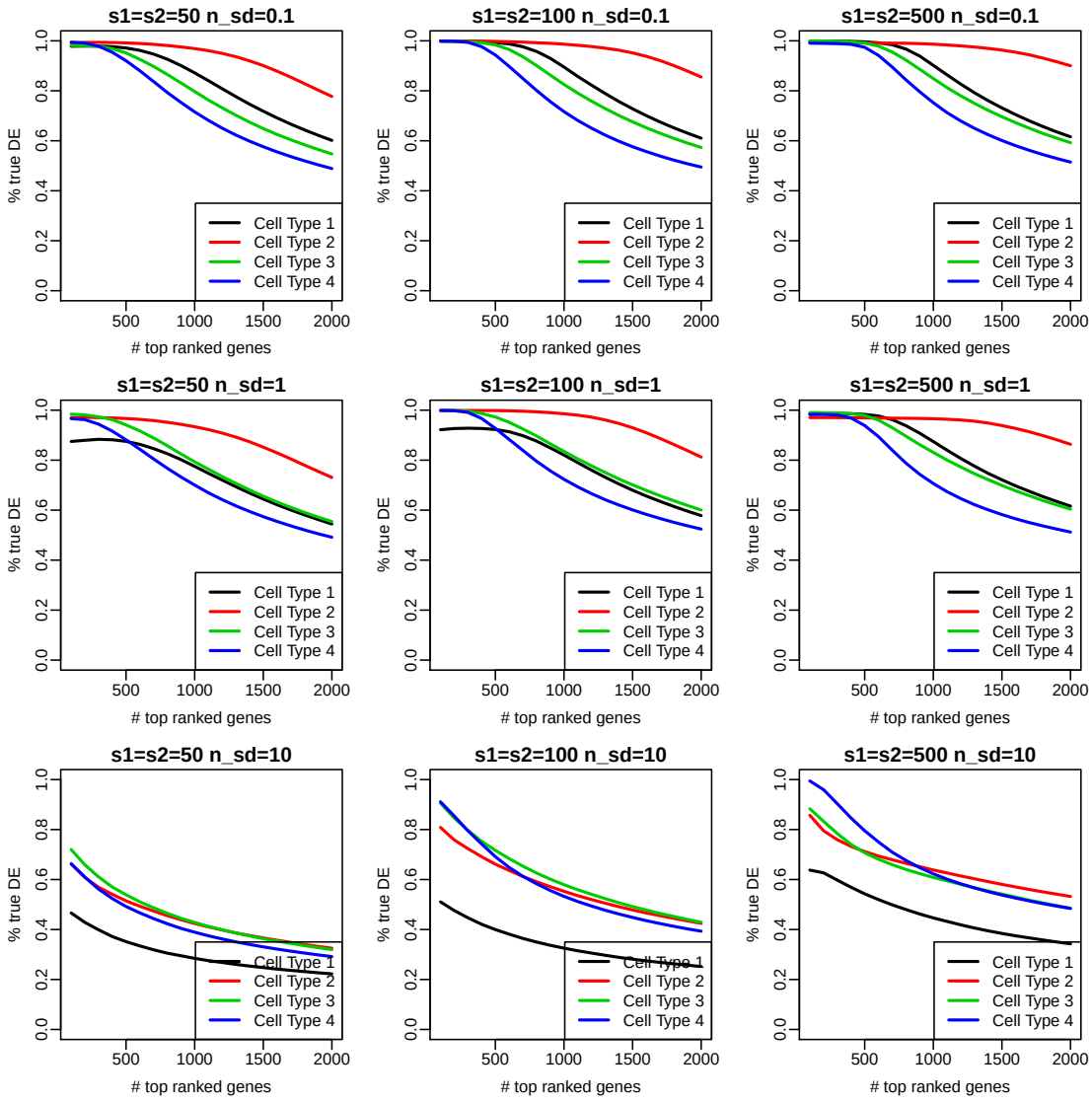


Figure S2: Impact of sample size and noise level on DE detection accuracy. True Discovery Rate (TDR) curves using the proposed method when reference-based method is used as up-stream deconvolution method. Sample size increases from left column to right column and observation noise increases from top row to bottom row.

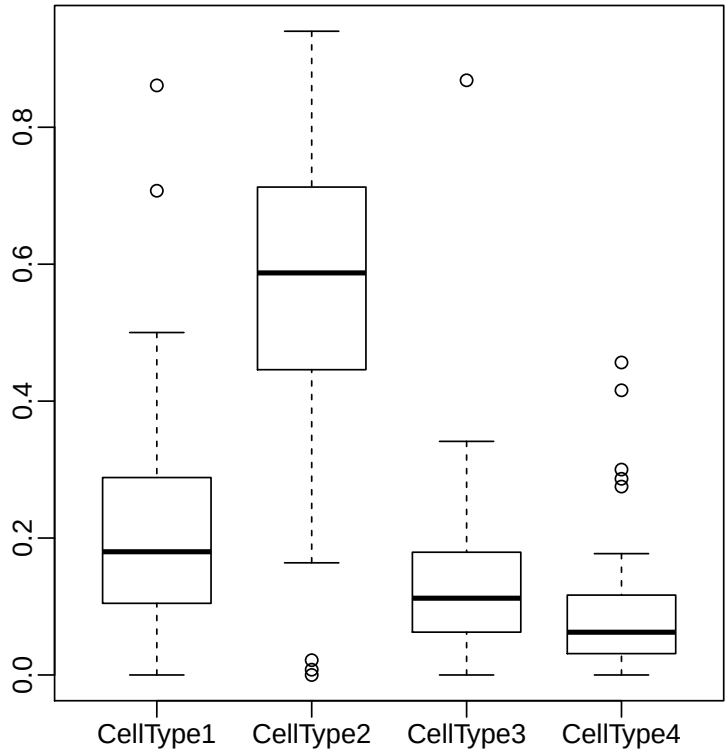


Figure S3: Boxplot of the proportion estimation of the AD proteomics dataset. Reference-free method deconf is applied to this dataset and the number of tissue types is set as 4.

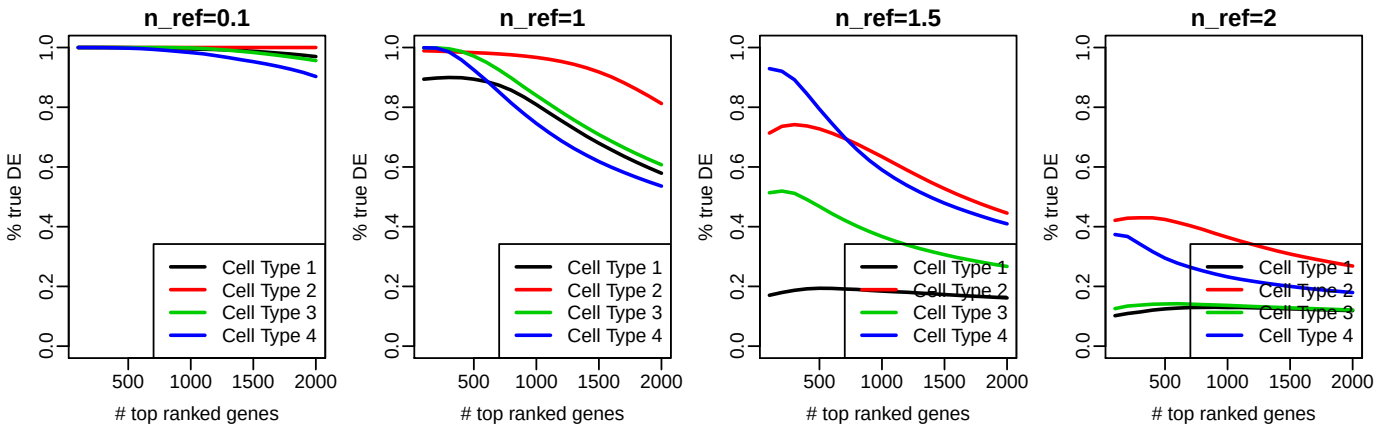


Figure S4: True Discovery Rate (TDR) plots with different levels of noise added to simulation reference panel. Reference-based method is used for deconvolution. The noise on pure tissue panel increases from left to right.

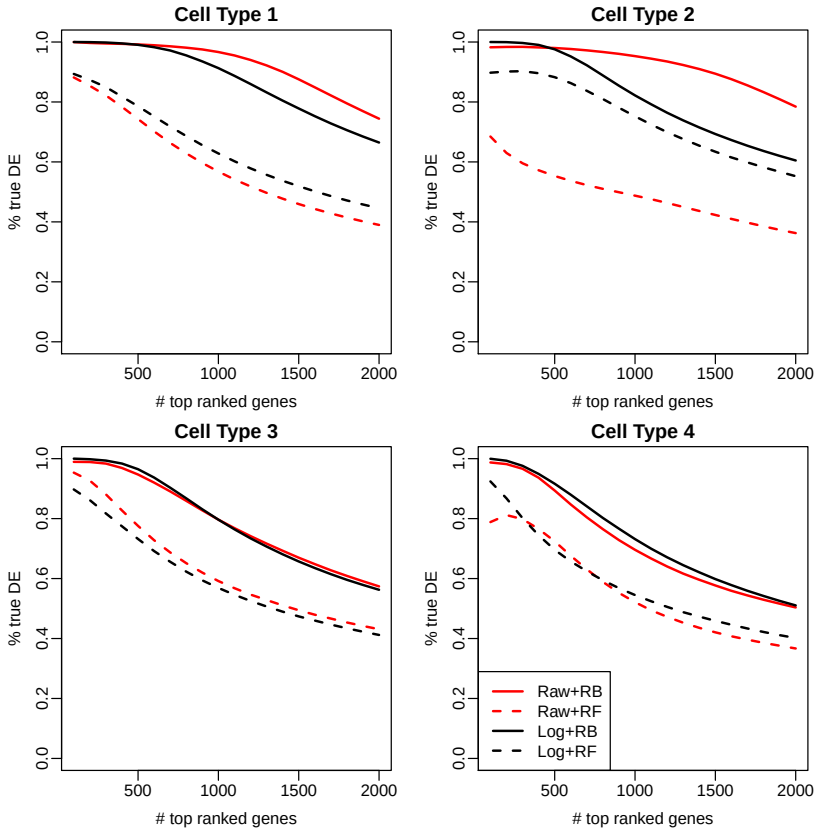


Figure S5: Impact of using raw vs. log-scale data in four cell types. TDR curvess from TOAST when using raw or log-scale data, and different proportion estimation as inputs for csDE detection. Raw: raw scale data. Log: log scale data.

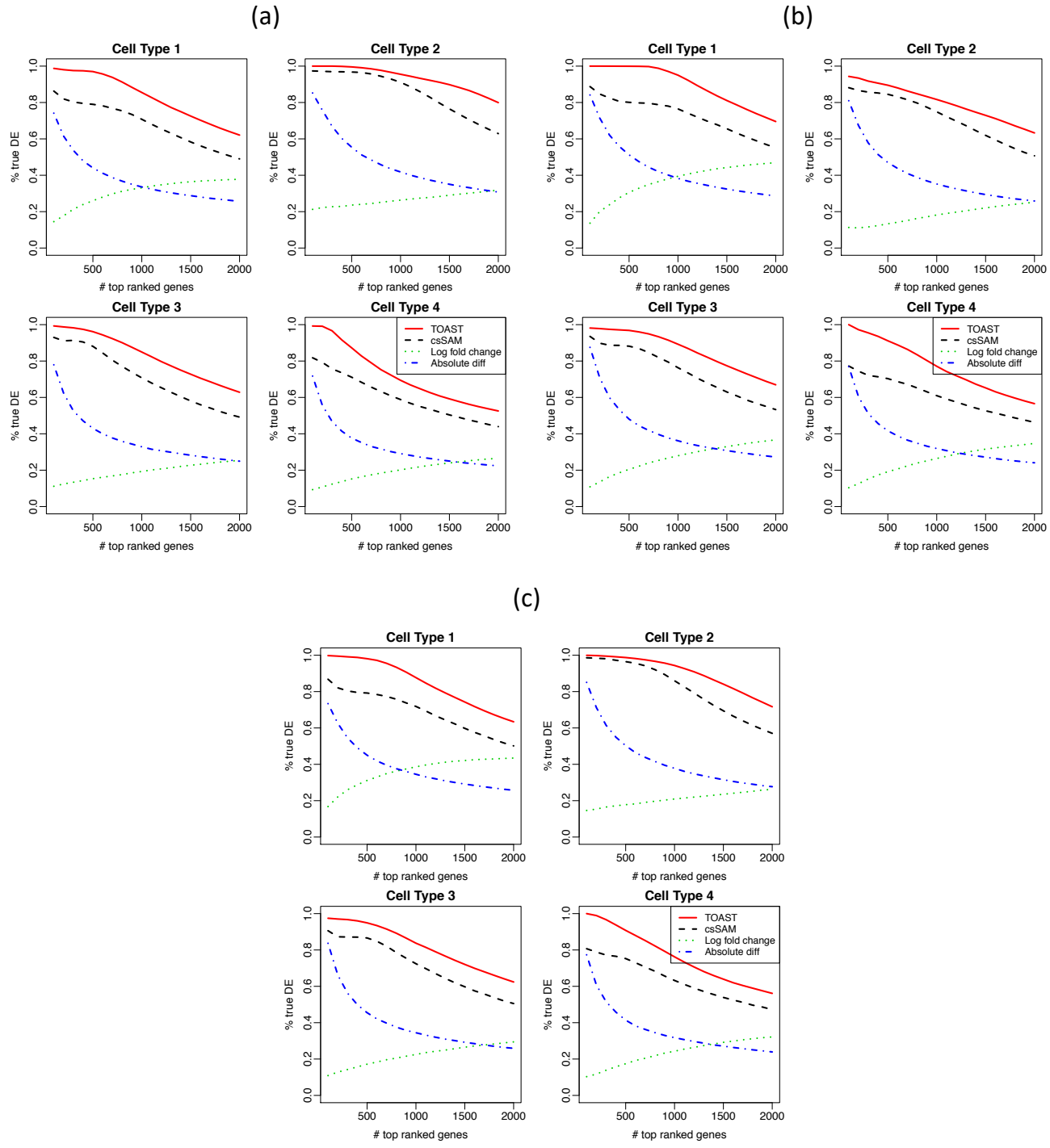


Figure S6: True Discovery Rate (TDR) plots of TOAST, csSAM, Log fold change and Absolute diff for all cell types in different DE generation settings. Figure (a): DE genes change in 2 cell types simultaneously. Figure (b): DE genes change in 3 cell types simultaneously. Figure (c): DE genes change in 4 cell types simultaneously. Reference-based method is used for deconvolution.

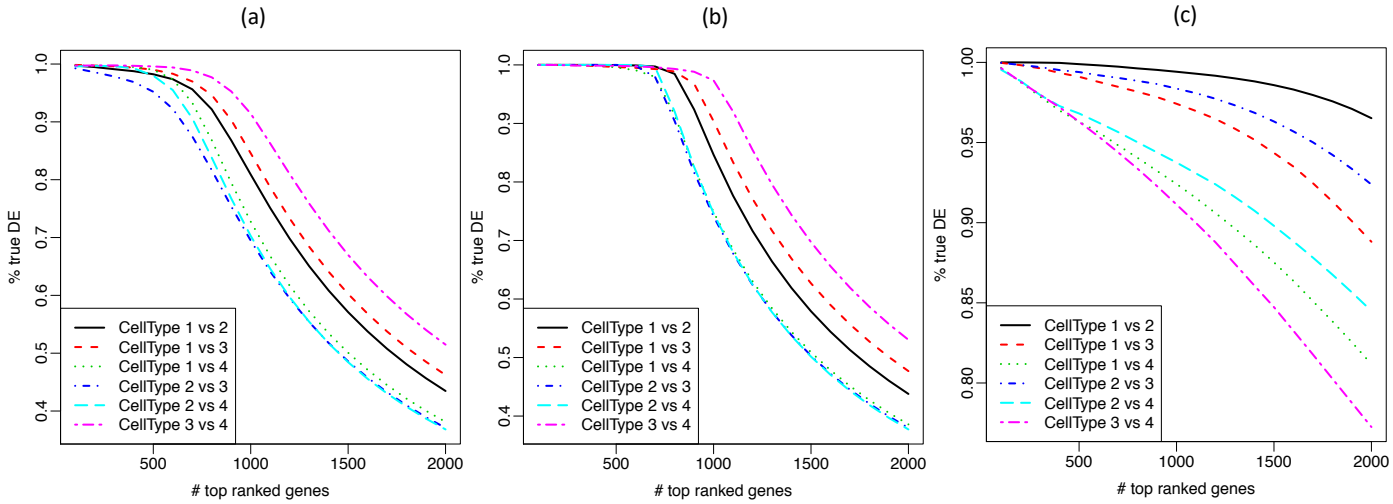


Figure S7: True Discovery Rate (TDR) plots of TOAST in testing for other hypothesis. Figure (a) is to test difference between cell types in normal group (Hypothesis 2 in Section 2.2); (b) is to test difference between cell types in case group (Hypothesis 3 in Section 2.2); (c) is to test difference of changes between cell types in two conditions (Hypothesis 4 in Section 2.2).

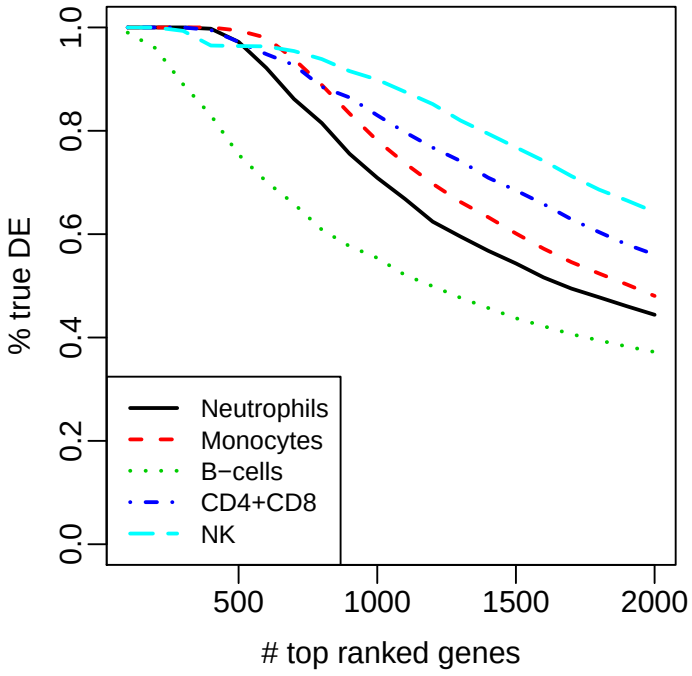


Figure S8: True Discovery Rate (TDR) plots of TOAST in the simulation study with RNA-seq dataset (GEO60424). Raw counts and true mixture proportions are used as inputs to TOAST.

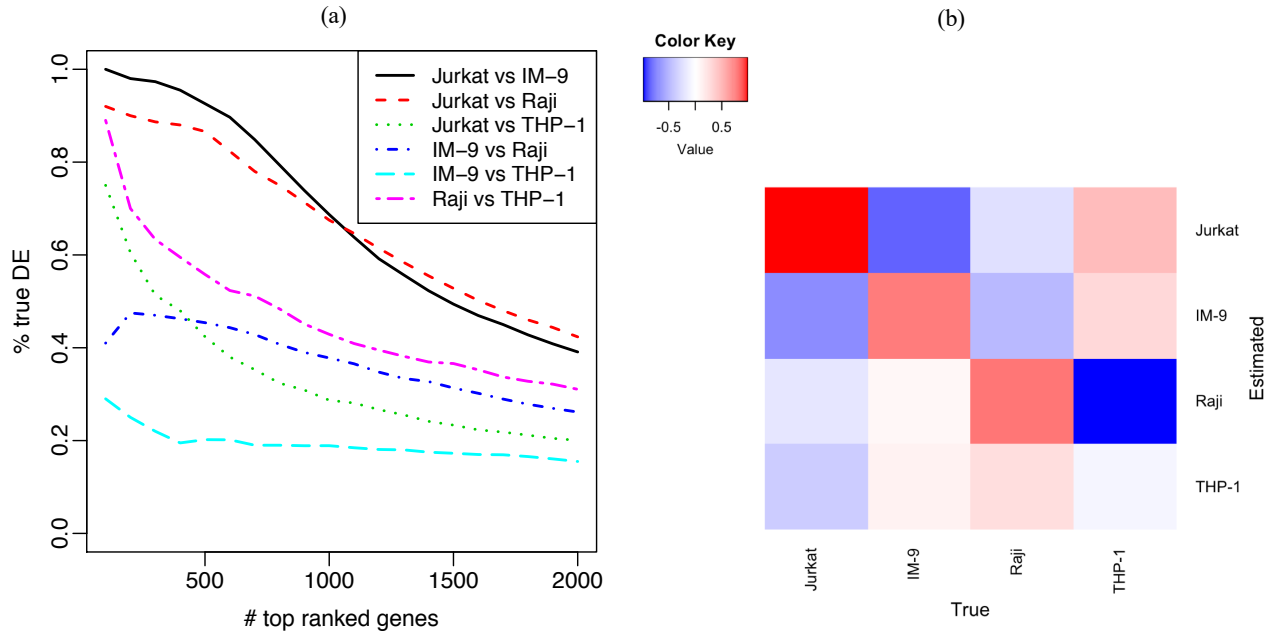


Figure S9: In the Immune data analysis, (a) TDR plot of across cell type DE detection using the proposed method when reference-free method is used for up-stream deconvolution analysis. (b) Heatmap of correlation coefficients for estimated proportions from reference-free methods versus true proportions.

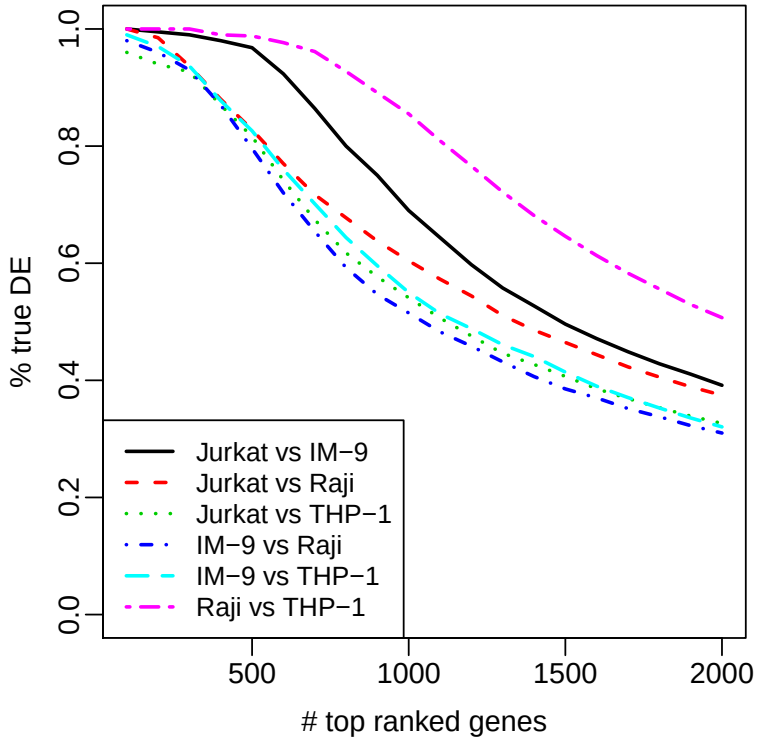


Figure S10: In the Immune data analysis, TDR plot of across cell type DE detection using the proposed method when true proportions are used as inputs.

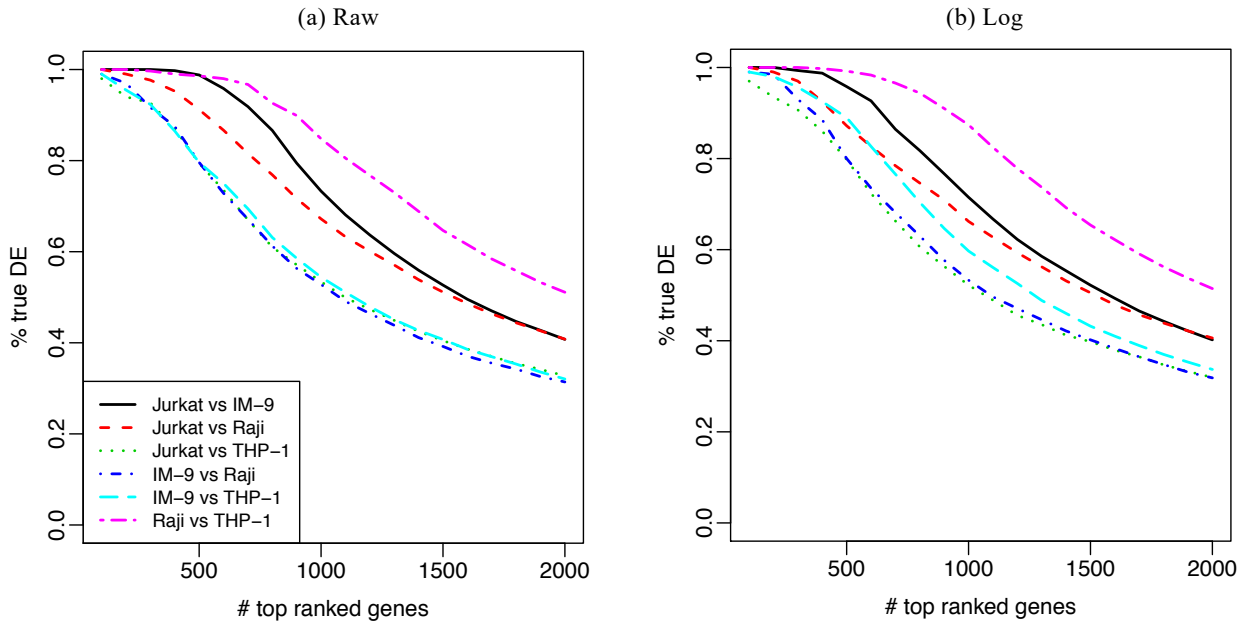
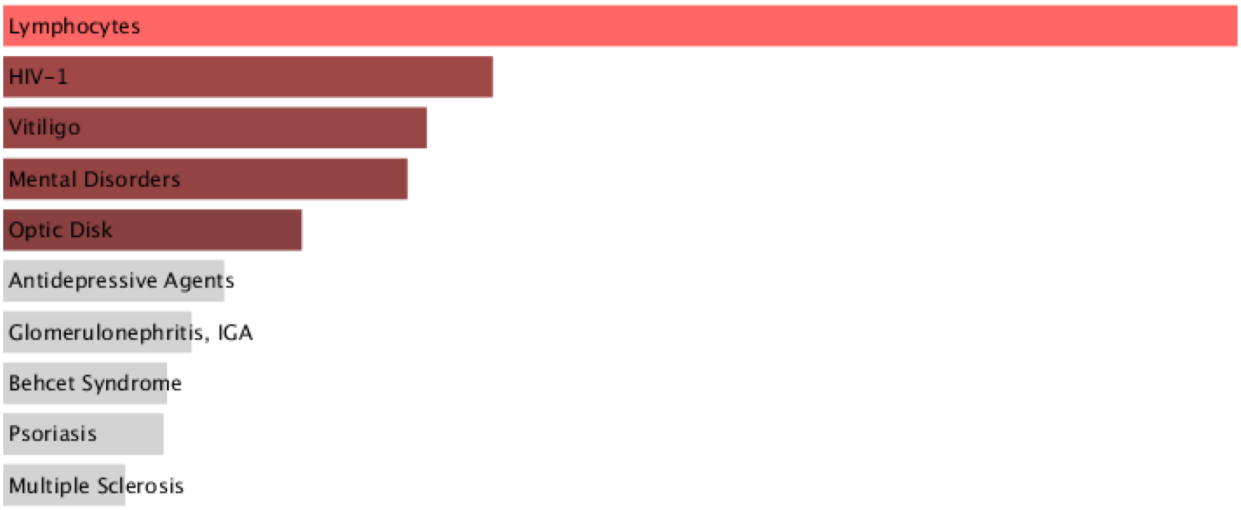


Figure S11: In the Immune data analysis, TDR plot of across cell type DE detection using the proposed method using raw-scale data or log-scale data. Reference-based deconvolution is used to obtain mixture proportions.

(a)



(b)

Index	Name	P-value	Adjusted p-value
1	Lymphocytes	0.0002109	0.02995
2	HIV-1	0.01212	0.6849
3	Vitiligo	0.01736	0.6849
4	Behcet Syndrome	0.07140	0.8419
5	Heart Failure	0.09765	0.8419
6	Psoriasis	0.07289	0.8419
7	Multiple Sclerosis	0.08948	0.8419
8	Antidepressive Agents	0.05238	0.8419
9	Fibrinogen	0.1064	0.8419
10	Optic Disk	0.03426	0.8419

Figure S12: Enrichment analysis of the database of Genotypes and Phenotypes (dbGaP) using EnrichR. Top panel (a) shows the enrichment level (log p value) of each term, bottom panel (b) shows the p value and adjusted p value of each term.