Distribution Agreement

Signature:

_______________________________ _______________

Qi Yu Date

Sub-tissue type eQTL analysis of GTEx data

By

Qi Yu

Master of Science in Public Health

Biostatistics and Bioinformatics

____

Zhaohui (Steve) Qin, PhD
Thesis Advisor

____

Hao Wu, PhD
Reader

Sub-tissue type eQTL analysis of GTEx data

By

Qi Yu

B.S.

Sun Yat-sen University

2018

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD

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
2020

Abstract

Sub-tissue type eQTL analysis of GTEx data

By Qi Yu

Background: Large-scale expression quantitative trait loci (eQTL) studies have been carried out recently to provide insights on how single-nucleotide polymorphisms associated with the expression of known genes. However, most of such studies ignored cell type mixing. Recent studies showed there are differences in gene expression patterns among different cell types. Thus tissue-specific eQTL studies with tissues of mixed cell type may suffer from false positives problems. In this work, we provide a new tool for sub-tissue type eQTL analysis with a recently published method (TOAST) for identifying cell-type specific effects.

Materials and Methods: Here we only consider the case of whole blood in GTEx project. We test both reference-based methods--CYBERSORTx and PRC with reference LM22 and a reference-free method--TOAST for deconvolution. We then conduct sub-tissue type specific eQTL analysis using TOAST.

Results: Deconvolution analysis show that there is a significant difference in the proportions of the six major cell types found in whole blood (CD4 T Cells, CD8 T Cells, B cells, Monocytes, Neutrophils and NK cells) across samples. Cell-type specific eQTL analysis on gene MARK4 with its significant associated SNPs in whole blood shows that eQTLs of MARK4 are more significant for Neutrophils than that for other cell types in the mixture.

Conclusion: We present a novel sub-tissue type eQTL analysis tool, can be applied to expression data measured from whole tissues provided knowledge of the reference cell types of the tissue is known. Our study reveals that eQTL analysis can be conducted at the sub-tissue type level when a reference is available.

Sub-tissue type eQTL analysis of GTEx data

By

Qi Yu

B.S.

Sun Yat-sen University

2018

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD

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
2020

ACKNOWLEDGEMENT

I would first thank my thesis advisor Dr. Steve Qin of Rollins School of Public Health at Emory University. He gave me a great help and encouragement in academic filed, answered my question patiently, taught me how to perform a proper research and steering me in the right direction. I would also like to thank my thesis reader Dr. Hao Wu for his great help during the revision process. Thank you!

1. Introduction

Thousands of genome-wide association studies (GWASs) have been published to date. Subsequently, large-scale expression quantitative trait loci (eQTL) datasets are studied to provide insights into how single-nucleotide polymorphisms (SNPs) associated with the expression of known genes. The latest Genotype-Tissue Expression (GTEx) project [12] (version 8) collects 54 human tissues from each of 948 donors and identifies eQTLs by associating genotypes called from whole-genome sequencing with gene expression levels obtained from bulk RNA-Seq.

The GTEx data, however, ignored cell type mixing. They only provide gene expression data with a weighted average of signals from multiple cell types for different tissues. The QTL studies with tissues of mixed cell type may suffer from additional challenges because the mixing proportions may be confounded with covariates (PEER factors) used in eQTL analysis. The confounding yields false positives in eQTL and co-expression analysis. [19]

There are several methods and software published for identifying cell-type specifics effects [11, 10]. The first step for these methods is figuring out cell mixture proportions. Existing cell-type proportion estimate methods mainly fall into two major categories: reference-based deconvolution [14, 5, 1, 4] and reference-free deconvolution [10, 16, 3, 18, 9, 2, 8]. The reference-based deconvolution is limited to known reference panels, where existing reference panels are only available in the brain [6], pancreas [13], and blood [22, 1, 17]. The reference-free deconvolution is applied when the reference panels are not available or there are mixed tissues. With known cell-type proportion, a linear based model framework can be applied to conducting the cell-type-specific analysis.[11]

In this work, we provide a new tool, for cell-type-specific eQTL analysis. This tool provides an RNA-seq deconvolution algorithm with both a reference-based and reference-free method, and a linear model-based framework for eQTL analysis. Compared to other eQTL analysis tools, our work can identify cell-type-specific eQTLs without knowing cell type profiles.

2. Methods and Materials

The data used for the analyses described in this manuscript were obtained from the GTEx Portal and dbGap accession number phs000424.v8.p2. The current release of GTEx (V8) includes 17,382 RNA-Seq samples from 948 donors. In this work, we focus on blood-tissue cis-QTL Data downloaded from GTEx Portal and corresponding genotype data from dbGap.

2.1 Observation Parameters

Our dataset includes 369 samples. For each sample, there are 20315 genes with expression level, whole genotype calls, and 65 covariates used in eQTL analysis, including genotyping principle components and PEER factors.

2.2 Deconvolution Algorithm

The first step for our method is to obtain mixing proportions. The mixing proportion can be computationally measured by several existing methods. In this manuscript, we use reference-based deconvolution methods CYBERSORTx [15], Robust Partial Correlations (PRC) [21], Constrained Projection (CP)[7] and a reference-free deconvolution method TOAST RefFreeEWAS [10]

The Reference-based method treats the gene-expression profile of any given sample as a linear combination of a given set of reference gene-expression profiles underlying some specific cell-types. Given a number C of underlying cell-types, each with a gene-expression profile b_c and denoting by y the gene-expression profile of a given sample, the underlying model is

$$y = \sum_{c=1}^{C} w_c b_c + \epsilon \tag{1}$$

The general idea is to estimate the weight coefficients for each cell-type in a least squares sense. Assuming that the reference database contains some specific cell-types presenting in the sample y, one may assume that $\sum_{c=1}^{C} w_c = 1$ (or more generally that $\sum_{c=1}^{C} w_c \leq 1$) The 3 algorithms differ in how the normalization constraint is implemented:

CYBERSORTx and PRC restrict the weight of the cell types that must be non-negative and add to one after the deconvolution process. CYBERSORTx enforces the constraints using a machine learning-based method and RPC applies robust multivariate linear regression and robust partial correlations.

CP performs the inference of weight least squarely but imposes the positivity and normalization constraints as a part of the inference process.

The reference-free method based on kinds of factor analysis. It selects a set of "informative" features that contains the information for cell proportion. TOAST provides an efficient feature selection procedure to improve RefFreeEWAS [8]. The key idea is to identify features showing distinct profiles across different cell types, without knowing the pure cell type profiles or mixing proportions a priori. The feature selection procedure is purely data-driven, without requiring any additional information. When the reference is unavailable, the reference-free method is the only solution.

2.3 Cell type-specific eQTL analysis

With a known mixing proportion, we conduct cell-type-specific eQTL analysis with the differential signals dissecting method proposed by Li (2018) [11]. In our work, we have G genes and N samples. Denote the observed expression level for gth gene and ith sample by Y_{gi} We assume there are K cell types in the mixture, and the proportions obtained for sample i are $\theta_i = (\theta_{i_1}, \theta_{i_2}, \dots, \theta_{i_K})$. Here we note that for reference-based method CYBERSORT, there are six types of purified blood cells (K=6, Nk, Bcell, CD8Tcell, CD4Tcell, Monocytes, Neutrophils). For reference-based method TOAST RefFreeEWAS, K can be set as a constant number, with unknown exact cell type.

For the gth gene in the ith sample, denote the unobserved expression in the kth cell type as $X_{gik.}$ For simplicity of notation, we will drop the subscript g in the following derivation. And our method will be performed one gene at a time in the same manner. Let δ_i be the genotype and Z_i be a vector for covariates, including genotyping principle components and PEER factors.

We assume the pure cell type profile satisfies: $E(X_{ik}) = u_{k\,+}\, \delta_i\rho_k + Z_i\beta_k$. Here u_k represents the mean level for cell type K, and $\rho_{k,}$ β_k are coefficients associated with the covariates. The observed data Y_i is the weighted average of X_{ik} 's. For sample i, given the proportion θ_i we have

$$E[Y_i;\theta_i] = \sum_k \theta_{ik} E[X_{ik}] = \sum_k (\theta_{ik}\mu_k + \theta_{ik}\delta_i\rho_k + \theta_{ik} Z_i^T \beta_k) \tag{2}$$

This is a linear based model, which includes mixing proportion as main effects and mixing proportion by covariate interactions. Assume we have Y from a total of N samples. Denote all observed data as $Y \;=\; [Y_1, Y_2, \cdots, Y_{\,N}]^T$, the observed data can be described as a linear model:

$$E(Y) = V\beta \tag{3}$$

Where:

$$V = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1K} & \theta_{11}\delta_1 & \cdots & \theta_{1K}\delta_1 & \theta_{11}Z_1^T & \cdots & \theta_{1K}Z_1^T \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2K} & \theta_{21}\delta_2 & \cdots & \theta_{2K}\delta_2 & \theta_{21}Z_2^T & \cdots & \theta_{2K}Z_2^T \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \theta_{N1} & \theta_{N2} & \cdots & \theta_{NK} & \theta_{N1}\delta_N & \cdots & \theta_{NK}\delta_N & \theta_{N1}Z_N^T & \cdots & \theta_{NK}Z_N^T \end{bmatrix} \tag{4}$$

$$\beta_p = \left[\mu_1, \mu_2, \cdots, \mu_K, \rho_1, \rho_2, \cdots, \rho_K, \beta_1^T, \beta_2^T, \cdots, \beta_K^T\right]^T \tag{5}$$

Using this model, we can conduct a cell-type-specific eQTL analysis. Note that genotype δ_i has three levels (0, 1, 2). Therefore, we can perform F-test on k cell-type-specific eQTL analysis by following hypothesis test:

$$H_0 : All\ \rho_i = 0,\ where\ i = 1...K \tag{6}$$

3. Results

3.1 Deconvolution

3.1.1 Reference based deconvolution CYBERSORTx on blood tissue

In this work, we only consider the case of blood tissue, in which the main constituent cell types are well known. We use the leukocyte gene signature matrix (LM22) designed by CYBERSORT team as a blood reference which contains 6 blood cell types (CD4.T.cells, CD.8.T.cells, B cells, Monocytes, Neutrophils, and NK.cells)

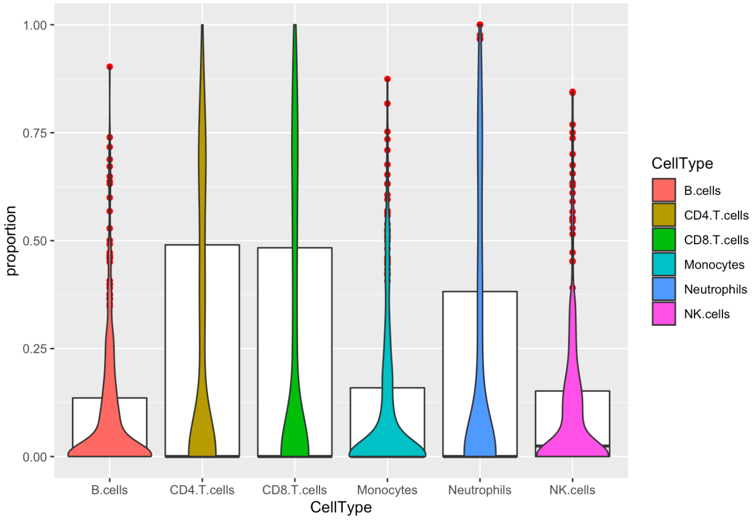


Figure 1, box plot of cell proportion in whole database

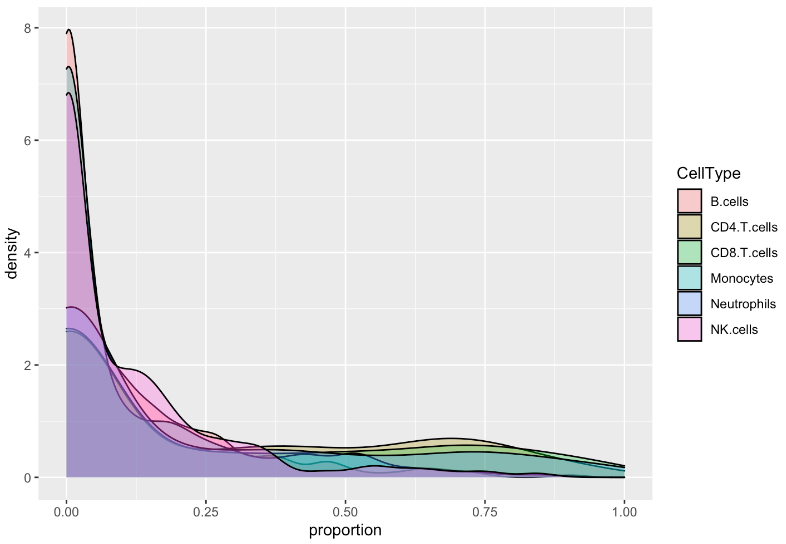


Figure 2, Density plot of cell proportion in whole database

Figure 1 and figure 2 show the proportion of six cell types in the whole database. The distribution of these proportions is all highly skewed, with a median around 0, but some high proportions for a few samples. Six cell types have different distribution patterns. CD4 T cells, CD8 T cells, and Neutrophils are three main components for most of the samples.

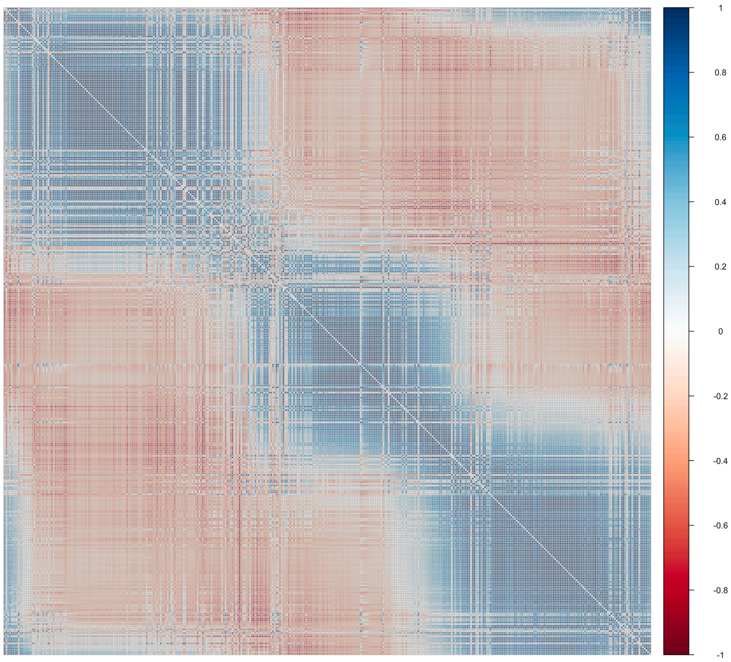
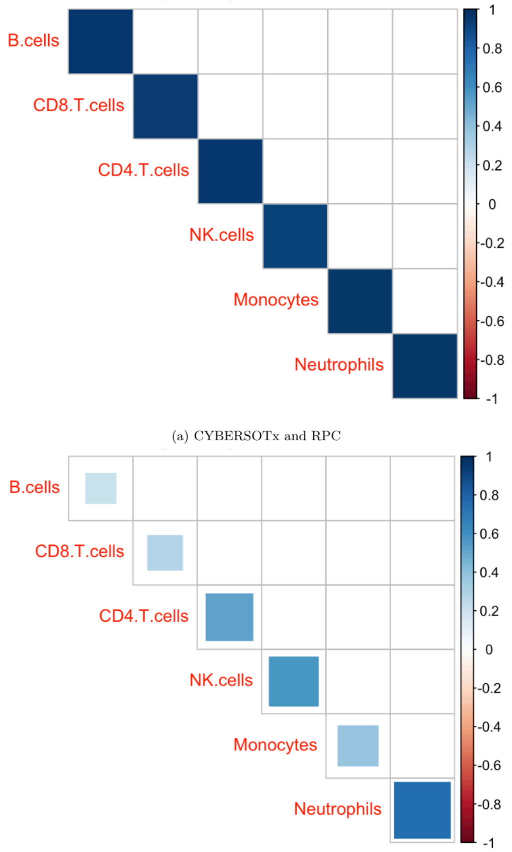


Figure 3, Pearson correlation coefficient matrices for 369 samples

Figure 3 shows the Pearson correlation coefficient matrices for 369 samples, indicating heterogeneity among samples. In the next section, we will compare eQTL analysis in a homozygous subgroup to that in a random group.

3.1.2 Comparison between reference based methods and reference free method.

We verify and compare the result in Figure 1 with alternative reference-based method RPC and reference-free method TOAST RefFreeEWAS. (Note: for this specific database and reference, reference-based method CP provided in our R package is unavailable due to inconsistent constraints in CP)



(a) CYBERSOTx and RPC

(b) CYBERSORTx and TOAST RefFreeEWAS

Figure 4, correlation plot of reference-based methods and reference-free method

Figure 4 shows the correlation plot for reference-free methods and reference-based methods with specific cell types. Results for reference-based methods CYBERSORTx and RPC are significantly correlated (with Pearson correlation coefficient 0.95), whereas results for reference-based method CYBERSORTx and reference-free method TOAST RefFreeEWAS are moderate positive correlated (with Pearson correlation coefficient 0.46).

It was reported that the reference-based method, in general, provides a more accurate and robust estimation than the reference-free method [14, 21, 23]. However, when reference is unavailable, the reference-free method is still a moderate reliable solution.

3.2 Cell-type specific eQTL analysis

In this manuscript, we pick gene MARK4 (ENSG00000007047.10) and its significant SNP pairs on chromosome 19 provided by the GTEx portal to illustrate our methods and findings. We use MARK4 due to a recent study that there is a functional requirement of MARK4 to maintain Neutrophils recruitment, and MARK4 expresses more significant in Neutrophils than in other cell types in blood.

3.2.1 Cell-type specific eQTL analysis on whole sample

We conduct a cell-type-specific eQTL analysis with the proposed method.

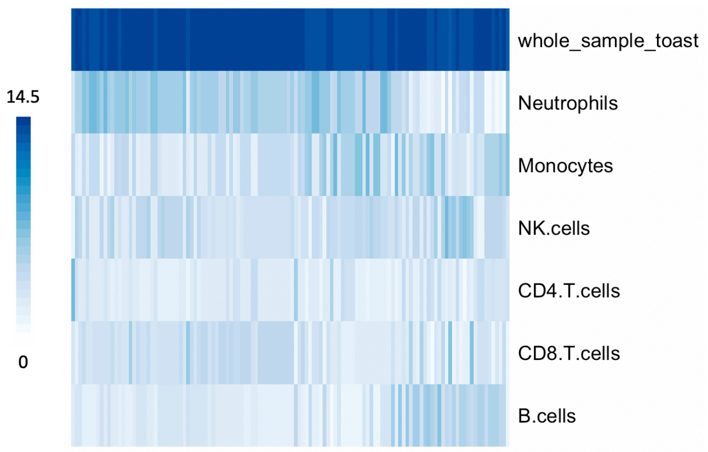


Figure 5, -log(p-value) matrices for whole sample eQTL and cell specific eQTL

Figure 5 shows that the whole sample produces more significant p-value for the same eQTL than that for any specific cell type. This difference due to a limitation of our method, we will consider improving our algorithm in future work.

To make this comparison between different cell types and the whole sample more straight forward. We only consider significant eQTLs with a p-value of less than 0.05. Our threshold 0.05 is larger than the standard in the GTEx portal. We choose 0.05 because of our small sample size and limited number of covariates in our database.

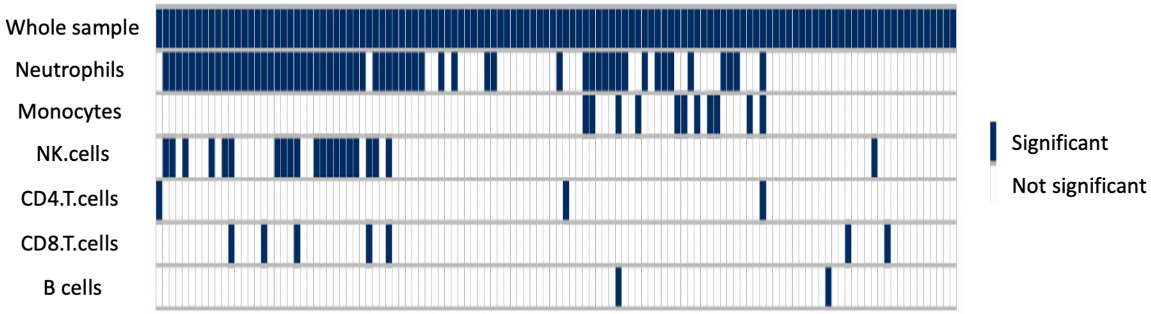


Figure 6, Significant whole sample eQTLs and cell-specific eQTLs

Cell type	Whole	Neutrophil	Monocytes	NK cells	CD 4 T cells	CD 8 T cells	B cells

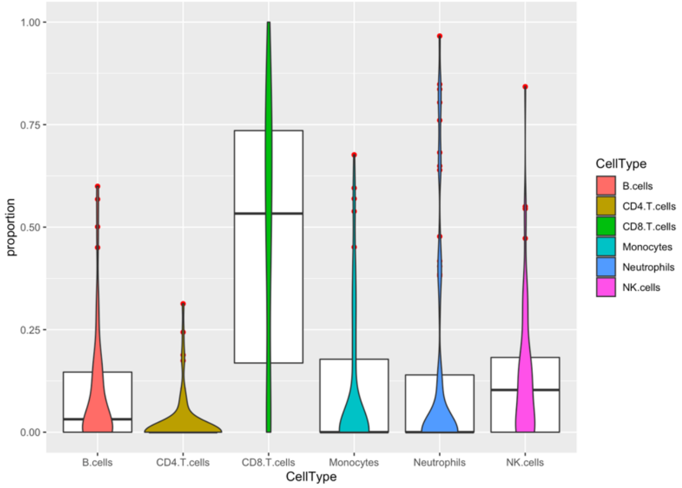
Num of sig eQTL	122	60	11	21	3	7	2

Table 1, Summary table for cell-specific eQTLs analysis

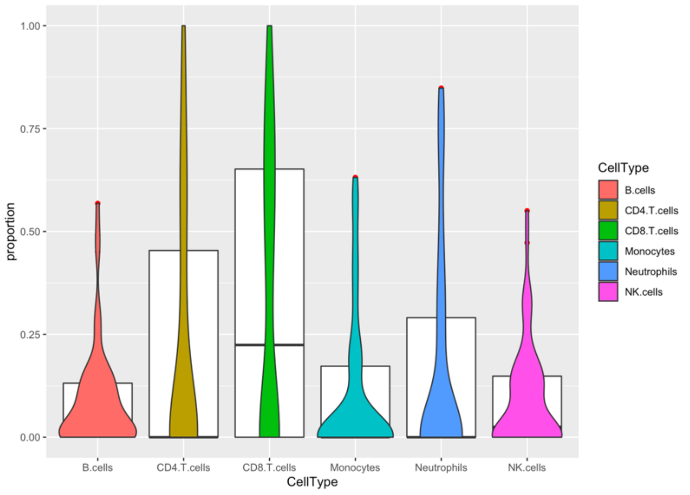
Figure 6 and table 1 show there are more significant eQTLs of MARK4 in Neutrophil than in other cell types, corresponding to the study MARK4 expresses more in Neutrophil than in other cell types. And for each cell type, it has specific eQTLs compared to other cell types. Therefore, our tool works well on finding cell-type-specific eQTLs.

3.2.2 Cell-type specific eQTLs in homozygous subgroup and random subgroup

Figure 4 shows there are three cluster groups in the whole sample. We conduct a cell-type-specific eQTL analysis on the first cluster and a random group with the same sample size.



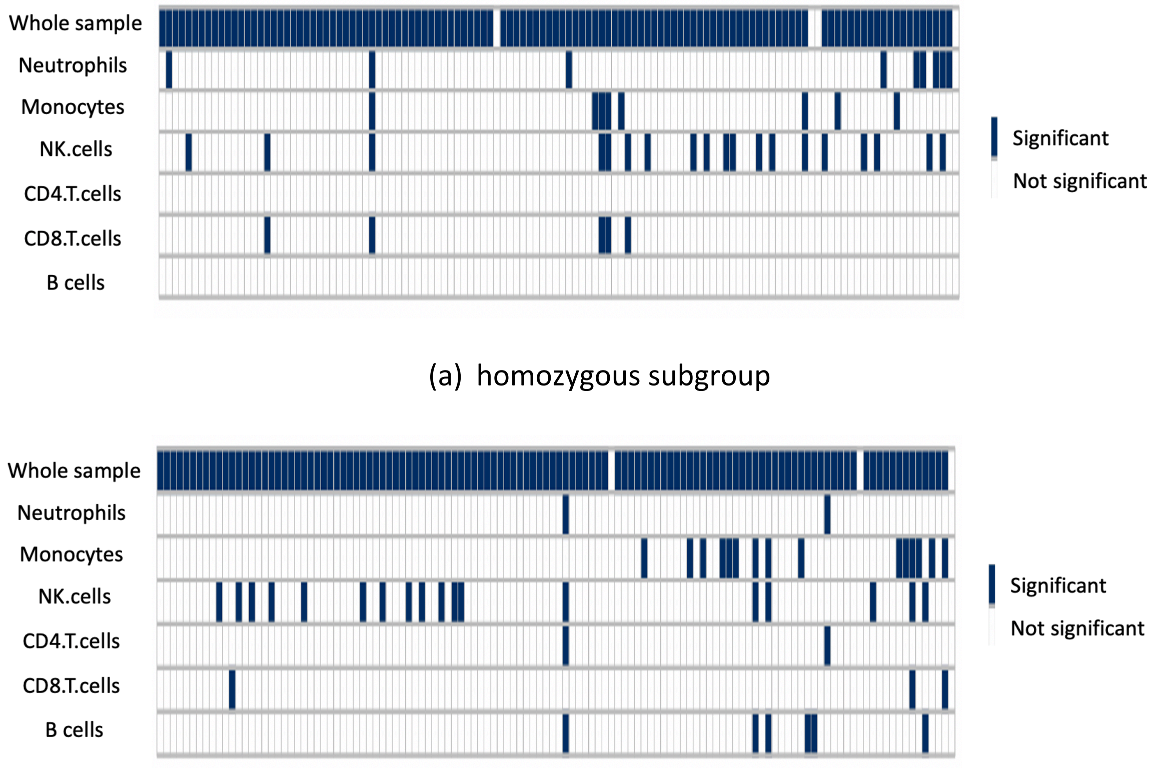
(a) homozygous subgroup



(b) random subgroup

Figure 7, box plot of cell proportion in homozygous subgroup and random subgroup

Figure 7 shows that homozygous subgroup has more B cells, CD 8 T cells and NK cells than random group.



(a) homozygous subgroup

(b) random subgroup

Figure 8, Significant cell-specific eQTLs in homozygous and random subgroups

Cell type	Whole	Neutrophil	Monocytes	NK cells	CD 4 T cells	CD 8 T cells	B cells
Num of sig eQTL(homozygous)	118	9	8	19	0	5	0
Num of sig eQTL(random)	119	2	15	18	2	3	6

Table 2, Summary table for cell-specific eQTLs in homozygous and random subgroups

Figure 8 and Table 2 indicate there isn't a specific pattern in a significant level change in homozygous and random subgroups. And compared to the whole database, Neutrophils do not have more significant eQTLs than other cell types. The possible reason is that our method does not perform well in a small database. Future work is needed to improve our tool.

4. Conclusion

In summary, we have presented a novel sub-tissue type eQTL analysis tool, for finding cell-specific eQTLs in mixtures. We expect wide applications of the proposed method on uncovering sub-tissue heterogeneity in eQTL analysis, and to this expectation, we are working on an R package which will be freely provided to the bioinformatics community.

5. Discussion

For the future version of our tool, we plan to consider several extensions and modifications. Firstly, we consider designing a fast and efficient cell-type-specific eQTL analysis tool which can perform testing on multiple transcript-SNP pair at a time. It may provide 5-10 orders of increase in performance. Secondly, we will consider using several methods dealing with missing data, rather than abandon the whole sample. Thirdly, we will consider an alternative method for testing how gene expression associated with genotypes. We will also consider designing a procedure to help users choose a better-performed deconvolution and method and make improvements on reference-free deconvolution.

6. References

[1]A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, and H. F. Clark,Deconvolution of blood microarray dataidentifies cellular activation patterns in systemic lupus erythematosus, PLOS ONE, 4 (2009), pp. 1–16.
[2]Y. Assenov, F. M ̈uller, P. Lutsik, J. Walter, T. Lengauer, and C. Bock,Comprehensive analysis of dna methylationdata with rnbeads, Nature Methods, 11 (2014), pp. 1138–1140.
[3]J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov,Metagenes and molecular pattern discovery using matrixfactorization, Proceedings of the National Academy of Sciences, 101 (2004), pp. 4164–4169.
[4]J. Clarke, P. Seo, and B. Clarke,Statistical expression deconvolution from mixed tissue samples, Bioinformatics, 26(2010), pp. 1043–1049.
[5]T. Gong, N. Hartmann, I. S. Kohane, V. Brinkmann, F. Staedtler, M. Letzkus, S. Bongiovanni, and J. D. Szus-takowski,Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complexclinical blood samples, PLOS ONE, 6 (2011), pp. 1–11.
[6]J. Guintivano, M. J. Aryee, and Z. A. Kaminsky,A cell epigenotype specific model for the correction of brain cellularheterogeneity bias and its application to age, brain region and major depression, Epigenetics, 8 (2013), pp. 290–302. PMID:23426267.
[7]E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke,and K. T. Kelsey,Dna methylation arrays as surrogate measures of cell mixture distribution, BMC Bioinformatics, 13(2012), p. 86.
[8]E. A. Houseman, M. L. Kile, D. C. Christiani, T. A. Ince, K. T. Kelsey, and C. J. Marsit,Reference-free deconvolutionof dna methylation data and mediation by cell composition effects, BMC Bioinformatics, 17 (2016), p. 259.
[9]E. A. Houseman, J. Molitor, and C. J. Marsit,Reference-free cell mixture adjustments in analysis of DNA methylationdata, Bioinformatics, 30 (2014), pp. 1431–1439
[10] Z. Li and H. Wu, Toast: improving reference-free cell composition estimation by cross-cell type differential analysis, Genome Biology, 20 (2019), p. 190.
[11]Z.Li,Z.Wu,P.Jin,andH.Wu,Dissectingdifferentialsignalsinhigh-throughputdatafromcomplextissues,Bioinformatics, 35 (2019), pp. 3898–3905.
[12] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Flem- ing, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson,

K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nico- lae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore, The genotype-tissue expression (gtex) project, Nature Genetics, 45 (2013), pp. 580–585.

[13] J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, K.-Y. Fu, E. Kiss, K. L. Spalding, G. Landesberg, A. Zick, A. Grinshpun, A. M. J. Shapiro, M. Grompe, A. D. Wittenberg, B. Glaser, R. Shemer, T. Kaplan, and Y. Dor, Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease, Nature Communications, 9 (2018), p. 5068.

[14] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles, Nature Methods, 12 (2015), pp. 453–457.

[15] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, and A. A. Alizadeh, Determining cell type abundance and expression from bulk tissues with digital cytometry, Nature Biotechnology, 37 (2019), pp. 773–782.

[16] E. Rahmani, R. Schweiger, L. Shenhav, T. Wingert, I. Hofer, E. Gabel, E. Eskin, and E. Halperin, Bayescce: a bayesian framework for estimating cell-type composition from dna methylation without the need for methylation reference, Genome Biology, 19 (2018), p. 141.

[17] L. E. Reinius, N. Acevedo, M. Joerink, G. Pershagen, S.-E. Dahln, D. Greco, C. Sderhll, A. Scheynius, and J. Kere, Differential dna methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility, PLOS ONE, 7 (2012), pp. 1–13.

[18] D. Repsilber, S. Kern, A. Telaar, G. Walzl, G. F. Black, J. Selbig, S. K. Parida, S. H. Kaufmann, and M. Jacobsen, Biomarker discovery in heterogeneous tissue samples - taking the in-silico deconfounding approach, BMC Bioinformatics, 11 (2010), p. 27.

[19] A. Saha and A. Battle, False positives in trans-eqtl and co-expression analyses arising from rna-sequencing alignment errors, F1000Research, 7 (2018), pp. 1860–1860. 30613398[pmid].

[20] A. A. Shabalin, Matrix eQTL: ultra fast eQTL analysis via large matrix operations, Bioinformatics, 28 (2012), pp. 1353– 1358.

[21] A. E. Teschendorff, C. E. Breeze, S. C. Zheng, and S. Beck, A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies, BMC Bioinformatics, 18 (2017), p. 105.
[22] F. Vallania, A. Tam, S. Lofgren, S. Schaffert, T. D. Azad, E. Bongen, W. Haynes, M. Alsup, M. Alonso, M. Davis, E. Engleman, and P. Khatri, Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accu- racy and reduces biological and technical biases, Nature Communications, 9 (2018), p. 4735.
[23] S. C. Zheng, S. Beck, A. E. Jaffe, D. C. Koestler, K. D. Hansen, A. E. Houseman, R. A. Irizarry, and A. E. Teschendorff, Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses, Nature Methods, 14 (2017), pp. 216–217.