Identification of cell-type-specific marker genes from co-expression patterns in tissue samples

Yixuan Qiu [1], Jiebiao Wang [2], Jing Lei [1], and Kathryn Roeder [1,3]

[1] Department of Statistics and Data Science, Carnegie Mellon University, USA
[2] Department of Biostatistics, University of Pittsburgh, USA
[3] Computational Biology Department, Carnegie Mellon University, USA

Abstract

Motivation: Marker genes, defined as genes that are expressed primarily in a single cell type, can be identified from the single cell transcriptome; however, such data are not always available for the many uses of marker genes, such as deconvolution of bulk tissue. Marker genes for a cell type, however, are highly correlated in bulk data, because their expression levels depend primarily on the proportion of that cell type in the samples. Therefore, when many tissue samples are analyzed, it is possible to identify these marker genes from the correlation pattern.

Results: To capitalize on this pattern, we develop a new algorithm to detect marker genes by combining published information about likely marker genes with bulk transcriptome data in the form of a semi-supervised algorithm. The algorithm then exploits the correlation structure of the bulk data to refine the published marker genes by adding or removing genes from the list.

Availability and implementation: We implement this method as an R package markerpen, hosted on https://github.com/yixuan/markerpen.

Contact: roeder@andrew.cmu.edu

1 Introduction

Cell-type-specific (CTS) genes, also known as marker genes, are genes that are highly expressed in one cell type, but lowly expressed in other types. These genes, which define cellular identity, are key to the analysis of RNA transcriptional data. Knowledge of marker genes gives insights into the core set of genes whose expression is shared among all cells of a given type, and will fill critical gaps in our understanding of cell biology and possibly the cellular origins of pathologies (Kelley et al., 2018). Marker genes are used to annotate cell clusters (Kiselev et al., 2017), to study cellular composition of bulk tissues (Oldham et al., 2008; Xu et al., 2013; Kelley et al., 2018; Luecken and Theis, 2019), to estimate cell type fraction via deconvolution (Gaujoux and Seoighe, 2012; Zhong et al., 2013; Abbas et al., 2009; Newman et al., 2015; Avila Cobos et al., 2018), and to estimate CTS expression directly from bulk tissue (Wang et al., 2020a,b).

Because marker genes are defined by their strong differential expression among cell types, a common approach to identifying them is to conduct statistical tests on CTS transcriptome data, typically single-cell RNA sequencing (RNA-seq). Genes that have significant expression differences between one specific cell type and all others are regarded as marker genes for this type (Kiselev et al., 2017). Despite the obvious appeal of this direct approach, the availability of CTS transcriptome data is a great challenge for many studies. The cost for single-cell sequencing is generally high, and in some cases, viable cells are hard to obtain for tissues like human brain. Even if public data sets are available, they might not correspond well with the data in hand, being collected at a different developmental period or a different functional portion of the organ. Furthermore, there is a trade-off between sequencing depth and the number of cells that can be analyzed, and for this reason the resulting single-cell transcriptome is quite noisy. An alternative way to obtain reference transcriptome data is to use single-cell RNA-seq data from another species (Zeisel et al., 2015); however, the quality of the obtained marker genes based on data from a different species is questionable. To this end, there is a need for a reliable statistical technique for detecting marker genes that does not require well matched single-cell RNA-seq data.

The objective of this inquiry is to develop a method for identifying a set of marker genes that describe the expression of the cells that constitute a tissue sample directly from the bulk transcriptome. We will take advantage of the conjecture that marker genes identifying a common cell type are highly correlated in samples of bulk transcriptome data, because their expression levels depend primarily on the proportion of that

cell class in each sample (Oldham et al., 2008; Kelley et al., 2018). Motivated by this insight, we develop a new algorithm called MarkerPen, short for **marker** gene detection via **pen**alized principal component analysis, to detect marker genes by combining prior marker information with bulk transcriptome data. MarkerPen is a semi-supervised algorithm that requires two pieces of information: a list of potential marker genes, typically obtained from the literature, past experience, or available single-cell RNA-seq data; and a bulk RNA-seq data set, viewed as a mixture of pure cells. The algorithm then exploits the bulk data to refine the published marker genes by adding and removing genes from the list.

In summary, MarkerPen is motivated by the following two key findings: (1) marker genes are statistically highly correlated under mild and sensible assumptions; (2) highly correlated genes can be detected by estimating the leading eigenvectors of the correlation matrix. We formulate the MarkerPen algorithm as a modified sparse principal component analysis (sparse PCA, Jolliffe et al., 2003; Zou et al., 2006; Zou and Xue, 2018), which simultaneously selects highly correlated genes and encodes prior information about markers into the model. Our simulation study and multiple data analyses of human brain transcriptomes demonstrate the superior performance of the proposed method.

2 Materials and methods

2.1 Related work

The MarkerPen algorithm follows the path of two pioneering publications, Xu et al. (2013) and Kelley et al. (2018), who noted that marker genes tend to be highly correlated in bulk tissue. MarkerPen solves the marker detection problem by making better use of bulk RNA-seq data. The motivation for these methods is straightforward: many tissues and subjects have been assessed for bulk tissue expression; the data tend to be of better quality; and collecting bulk data is less costly. Although bulk data alone do not provide CTS transcriptome information, they can be combined with prior knowledge of marker genes to improve the quality of published markers. For example, Xu et al. (2013) first obtained CTS genes in mouse brain as potential markers for human brain, and then performed co-expression network analysis on human brain bulk data to select highly correlated genes of each type as the refined marker genes. This method has shown good empirical results, but has the drawback that genes can only be removed from the

candidate list, but not added from the complementary set. More recently, Kelley et al. (2018) applied a similar approach to the human brain transcriptome. They first built an unsupervised co-expression network for all genes, and then identified gene clusters that were maximally enriched with published markers. Each gene was then assigned a fidelity score for each cell type, as an indicator for the strength of association between the gene and the cell type. These scores, however, were based on the aggregation of multiple data sets, and hence the selected marker sets may be suboptimal for a specific study.

Both methods described above assume that marker genes tend to be highly correlated, which is an intuitive assumption supported empirically in numerous species (Oldham et al., 2008; Fertuzinhos et al., 2014; Ponomarev et al., 2010; Hilliard et al., 2012; Bakken et al., 2016; Hawrylycz et al., 2015), but lacks rigorous statistical justification. To resolve this shortcoming, in the supplementary material (Section S.1) we explicitly study the statistical properties of marker genes, and show that under weak assumptions the marker genes for the same cell type are highly correlated in the bulk data. Given this fact, we are then able to utilize the correlation structure to detect marker genes via the MarkerPen algorithm.

2.2 The MarkerPen algorithm

Because high mutual correlation is a necessary condition for marker genes, the first step of marker gene selection is to find a subset from the whole genome such that genes in this set are highly correlated with each other. If the true correlation matrix Σ is available, then such a goal can be achieved by computing PCA on Σ, as the eigenvectors of Σ, also known as factor loadings, indicate the contribution of each gene to form a gene group. In the case of a marker gene group, the eigenvector contains a few strong signals and a large number of small values, where the large coefficients correspond to highly correlated genes (Section S.2, Figure S1).

However, in practice, only the sample correlation matrix S is given, and S can be of very high dimension. Theoretical results show that conventional PCA is likely to fail in high dimensions (Johnstone and Lu, 2009; Jung and Marron, 2009), so in this case the sparse PCA method is preferred, which directly estimates a sparse eigenvector, meaning that most entries in this vector are zeros. Sparse PCA has many different variants, and in this article we consider the Fantope projection and selection algorithm (FPS, Vu et al., 2013), because it solves a convex optimization problem that has a

global convergence guarantee. Let $\Gamma_{p\times d}$ denote the eigenvectors of Σ associated with the largest d eigenvalues, where p is the number of genes, and then FPS estimates the top-d projection matrix $\Pi_{p\times p} = \Gamma\Gamma^{\mathrm{T}}$ by solving

$$\begin{aligned} \max_{X} \quad & \mathrm{tr}(SX) - \lambda\|X\|_{1,1} \\ \text{s.t.} \quad & O \preceq X \preceq I \text{ and } \mathrm{tr}(X) = d, \end{aligned}$$

where $\mathrm{tr}(A)$ is the trace of a matrix A, $\|X\|_{1,1} = \sum_{i,j}|X_{ij}|$ is the sum of absolute values of the elements in X, λ is a tuning parameter that controls the sparsity of eigenvectors, and $O \preceq X \preceq I$ means all eigenvalues of X are between 0 and 1. Once we get an estimate $\hat{\Pi}$ for the projection matrix Π, we can recover the eigenvectors Γ by computing the eigen decomposition of $\hat{\Pi}$, $\hat{\Pi} = \hat{\Gamma}\hat{D}\hat{\Gamma}^{\mathrm{T}}$.

In practice, there is abundant prior information about the marker gene list in the literature, which provides useful knowledge about the relationship between cell types and genes; however, such information is not exploited by FPS, resulting in low utilization of the available information. To fix this issue, the proposed MarkerPen algorithm modifies the original FPS such that prior information about markers can be combined with the collected bulk data. For simplicity, we first consider the detection of marker genes for one cell type. Let G be the indices of published marker genes for a cell type C, and then we solve

$$\begin{aligned} \max_{X} \quad & \mathrm{tr}(SX) - \lambda p_{G,w}(X) \\ \text{s.t.} \quad & O \preceq X \preceq I,\ X \geq 0,\ \text{and } \mathrm{tr}(X) = 1, \end{aligned} \tag{1}$$

to estimate the projection matrix $\Pi = \gamma\gamma^{\mathrm{T}}$, where γ is the leading eigenvector, $p_{G,w}(X) = \sum_{i,j}\tilde{p}_{G,w}(X_{ij})$ is a penalty function defined as

$$\tilde{p}_{G,w}(X_{ij}) = \begin{cases} |X_{ij}|, & i,j \in G \\ w^2|X_{ij}|, & i \notin G, j \notin G \\ w|X_{ij}|, & \text{otherwise} \end{cases},$$

and $X \geq 0$ means all elements of X are nonnegative. The added constraint $X \geq 0$ is based on the fact that marker genes are positively correlated, so both the eigenvector γ and the projection matrix Π have nonnegative entries. The extra tuning parameter

$w \geq 1$ is used to put larger sparsity penalty on genes that are not in the prior list G, so that genes outside G are less likely to be selected as marker genes, unless they show large signals. The optimization problem (1) can be solved via the proximal-proximal-gradient method (Ryu and Yin, 2017), with details in the supplementary material (Section S.3).

After we obtain the estimate for the leading eigenvector γ, we select genes that have coefficients greater than some threshold $\varepsilon > 0$, and treat them as marker genes for cell type C. For multiple cell types $C_1, C_2, \ldots$, we repeatedly apply the algorithm above to compute different marker gene groups sequentially.

2.3 Data sources

In the next section we validate the performance of MarkerPen using a broad range of bulk and single-cell RNA-seq data, and here we provide some basic information of each data set. Below are the bulk tissue data used in this article:

1. **MSBB** The Mount Sinai/JJ Peters VA Medical Center Brain Bank cohort (Wang et al., 2018) contains RNA-seq data from human temporal cortex, with 425 control samples and 425 samples from patients with Alzheimer's disease (AD, Braak score ≥ 4). Only the control samples are used.

2. **ROSMAP** The Religious Orders Study and the Rush Memory and Aging Project (Mostafavi et al., 2018; De Jager et al., 2018) collects RNA-seq data from the human dorsolateral prefrontal cortex (DLPFC), with 288 control samples and 348 AD samples. Only the control samples are used.

3. **MayoRNAseq** The Mayo Clinic RNA-seq data set (Allen et al., 2016; Allen et al., 2018) contains human temporal cortex RNA-seq data with 28 control samples and 82 AD samples. Only the control samples are used.

4. **BrainVar** The BrainVar data set (Werling et al., 2020) consists of 176 samples from the human DLPFC across development, from 6 post-conception weeks to young adulthood. To be comparable with other data sets we exclude pre-natal brains and focus on subjects that are at least 6 months old (epoch 3), finally with a sample size of 45.

5. **CMC** The human brain RNA-seq data collected by the CommonMind Consortium (Fromer et al., 2016) contain 258 adult schizophrenia subjects and 279

adult control subjects, and only the control samples are used. As the original data set spans a broad range of ages, we further split the control group into two subsets, resulting in groups with ages less than or equal to 70 (sample size 164) and greater than 70 (sample size 115).

We also use single-cell and single-nucleus RNA-seq data sets:

1. Mathys et al. (2019) provides single-nucleus transcriptomes from DLPFC of 48 subjects with varying degrees of AD pathology. Only the data from 17 control subjects are used.

2. Darmanis et al. (2015) obtains single-cell RNA-seq data of human cortical tissues from eight adults and four embryonic samples. Only the adult data are used.

3. Li et al. (2018) collects single-nucleus RNA-seq data from DLPFC of three adult brains.

4. Zeisel et al. (2015) provides mouse cerebral cortex single-cell RNA-seq data.

3 Results

3.1 Quality of selected markers

In this section we demonstrate the quality of marker genes selected by MarkerPen from three different angles.

First, as explained in Section 2.1, we expect to see that marker genes for the same cell type are highly correlated in the bulk data. Therefore, the quality of selected marker genes can be visually examined by the correlation matrix. We study human brain bulk tissue RNA-seq data, and use the MSBB data set for illustration. To apply the MarkerPen algorithm, the prior marker gene list is obtained from existing literature, including 184 marker genes for astrocytes, 130 genes for oligodendrocytes, 319 genes for neurons (all three from Cahoy et al., 2008), 100 genes for microglia (Hickman et al., 2013), and 237 genes for endothelial cells (Butler et al., 2016). Figure 1A shows the sample correlation matrix of the published marker genes in the MSBB bulk data. It can be seen that the correlation matrix roughly forms five blocks, but the boundary between the blocks is not very clear as much noise exists.

Then we apply the MarkerPen algorithm to refine the given marker gene list. For each cell type, we restrict the search range to the union of the published marker genes

and the top 500 genes that have the highest fidelity scores given by Kelley et al. (2018). Figure 1C demonstrates the sample correlation matrix of the refined genes, in which 50 genes are selected for each cell type. It is clear that after the refinement, genes in the same block have much stronger mutual correlation, whereas genes in different blocks are only weakly correlated. In other words, genes refined by MarkerPen have a correlation structure that better fits the property of marker genes.

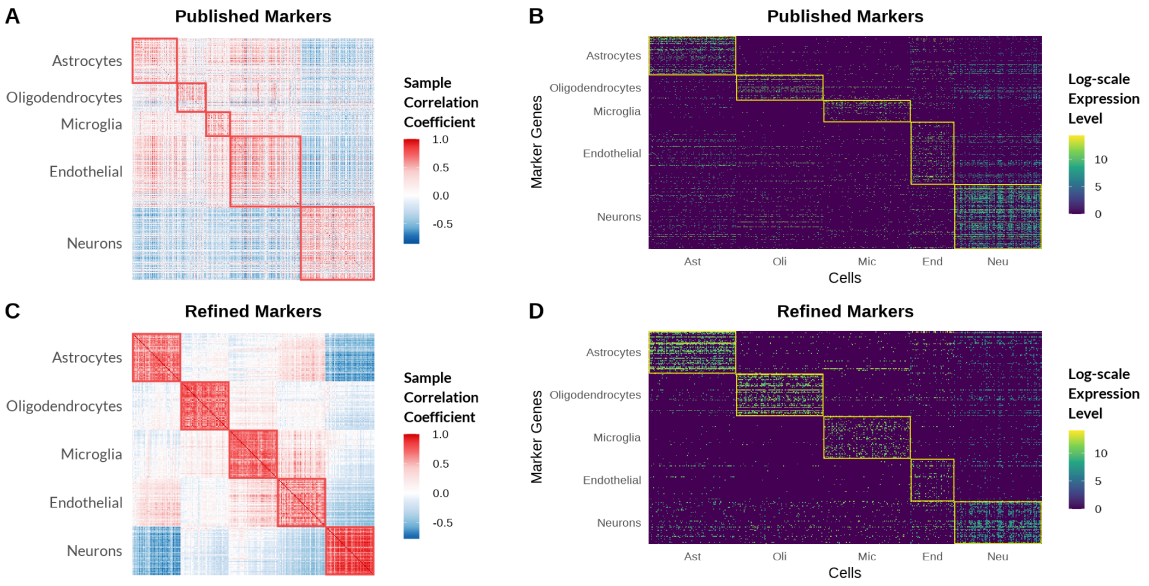

Figure 1: (A) Sample correlation matrix of published marker genes in the MSBB bulk data. (B) Gene expression of single-cell reference data from Mathys et al. (2019) on published marker genes. (C) Sample correlation matrix of refined marker genes output by MarkerPen. (D) Gene expression of single-cell reference data on refined marker genes.

Second, by definition, marker genes should be largely expressed in one cell type but weakly expressed in others. Therefore, it is helpful to examine the expression level of selected marker genes in purified single-cell data. We use the single-nucleus transcriptome data from Mathys et al. (2019) to demonstrate this idea. For each cell type, we randomly select 100 samples (50 for endothelial due to the limited number in the data set), and plot the logarithm-scale expression matrix on published and refined marker genes in Figure 1B and D, respectively. In Figure 1B, we can observe that many genes in the published list behave like noise, as they show very low expression level in virtually all cell types. In contrast, this defect has been greatly reduced in Figure 1D, where most noise genes have been removed by MarkerPen. This finding further justifies

the MarkerPen selection algorithm.

Finally, considering that the transcriptome data from Mathys et al. (2019) and the MSBB bulk data may not fully match, it is more appropriate to study the purified cells from the same subjects as in the bulk data. However in practice, this is not always possible. Instead, we can use the bMIND algorithm (Wang et al., 2020b) to estimate CTS gene expression for each subject in the bulk data. The output of bMIND can be viewed as the average of denoised single-cell data for the subjects in the bulk data. We plot the estimated CTS gene expression matrix on three types of markers: the published marker genes, the markers selected by MarkerPen, and the bMIND markers that are directly selected from the estimated CTS gene expression. The bMIND markers are treated as the ground truth. However, because marker genes form a highly correlated set, there is not a unique set of optimal genes to serve this purpose. In our evaluation we look to see if the set of selected markers matches the good properties exhibited in the bMIND set. The first row of Figure 2 demonstrates the results for the MSBB data set, from which we can find that published markers contain a lot of noise, whereas the MarkerPen output is very similar to that of bMIND. Also included in Figure 2 are the results for two additional bulk data sets: the ROSMAP and MayoRNAseq data. They both give similar results that validate the quality of MarkerPen genes.

3.2 Performance in downstream analysis

As marker genes are essential tools for many downstream analyses such as cell type fraction deconvolution, in this section we use simulation experiments to evaluate the performance of our algorithm in such tasks. Cell type fraction deconvolution is a problem commonly seen in bulk RNA-seq data analysis. Because the deconvolution result depends on the selection of marker genes, the quality of the selected markers can be measured by the estimation error of cell type fractions. We design a simulation experiment to compare MarkerPen with two supervised marker gene selection algorithms, with experiment setting described in the supplementary material (Section S.4, Figure S2, S3).

In practice, deconvolution can be conducted with or without single-cell reference samples, and the quality of reference samples may also vary. To reflect these different scenarios, we design three models for simulating the observed data:

1. **Matched reference case** Reference samples and the bulk data are simulated from the same signature matrix.

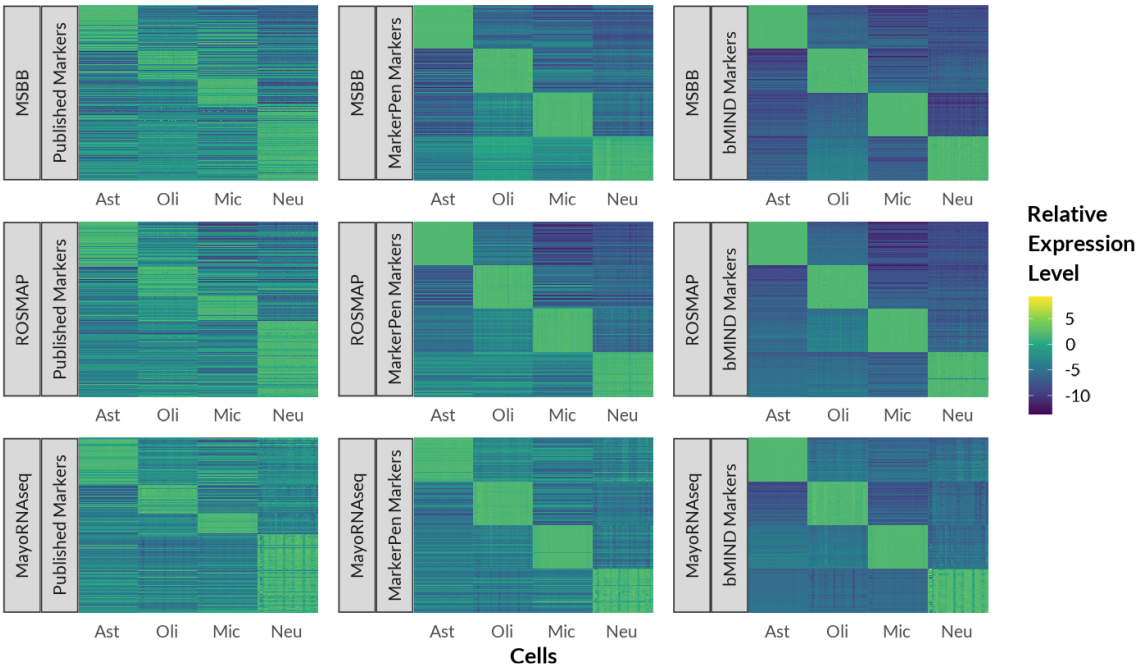

Figure 2: CTS gene expression of the MSBB, ROSMAP, and MayoRNAseq data sets on three types of marker genes: the published markers, the ones selected by MarkerPen, and the bMIND markers that can be treated as the truth. Ast=astrocytes, Oli=oligodendrocytes, Mic=microglia, Neu=neurons.

2. **Noisy reference case** The bulk data use a perturbed version of the signature matrix: some percentage of the genes, ranging from 5% to 30%, are set to noise. This indicates that some genes may be markers in the reference data, but they play no role in the bulk data.

3. **No reference case** No reference samples are simulated.

For model 1 and model 2, both the bulk data and the reference samples are available, and we use a supervised method, **dtangle** (Hunt et al., 2018), to accomplish the deconvolution. For model 3, only the bulk data and the marker gene list are available, so we apply a semi-supervised algorithm for deconvolution, the digital sorting algorithm (**DSA**, Zhong et al., 2013). The choice of deconvolution algorithms is beyond the scope of this article, as the main purpose of this section is to evaluate the effect of marker gene selection for a fixed deconvolution method. In practice any deconvolution algorithm that needs marker genes can be used in place of the methods investigated here.

In our experiments, we use the mouse brain single-cell RNA-seq data from Zeisel et al. (2015) to simulate the true signature matrix. We select seven major cell types (astrocytes, oligodendrocytes, microglia, endothelial, interneurons, S1 pyramidal neurons, and CA1 pyramidal neurons) from the whole single-cell data, and restrict to 2452 genes that are known to be associated with the cell types (Table S1 of Zeisel et al., 2015). Following the steps in Section S.4, we simulate the fraction matrix, reference samples, and the bulk data according to a stochastic model. From the signature matrix, we randomly select 50 genes from each cell type block, and treat them as known marker genes. Of course, due to the possible perturbation of the signature matrix, some of the claimed marker genes will be noise in the bulk data, and hence provide little information about the cell type. This treatment is used to mimic the quality of marker genes in reality.

We repeat the procedure above 30 times, so that in every simulation run, the generated data are different but follow the same stochastic model. We compute the deconvolution estimation errors in each simulation run, and summarize their distribution density curves in Figure 3.

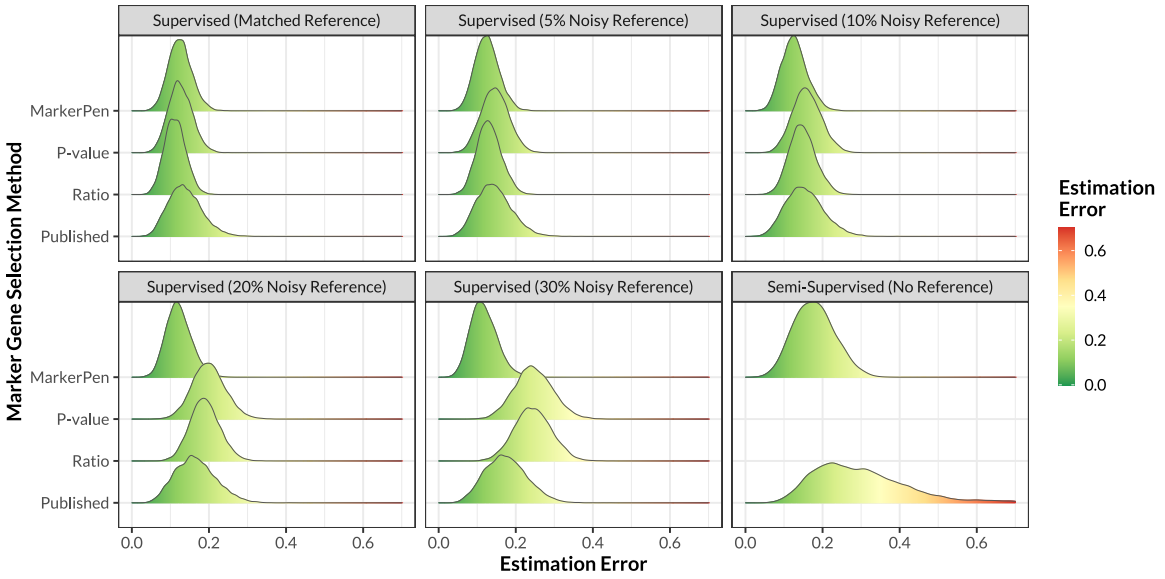

Figure 3: Impact of marker gene selection algorithm on deconvolution estimation error, displayed as distribution density curves. The vertical axis stands for different marker gene selection methods. MarkerPen: the proposed method. P-value and Ratio: selection methods based on reference samples, implemented in the **dtangle** R package. Published: using all published marker genes without selection.

In Figure 3, each panel represents one model for the reference sample. It is clear that when the reference sample and bulk data are matched, all marker gene selection methods behave equally well, compared with the last row that stands for no selection. However, when the noise level increases, selection methods purely based on the reference sample become much worse, whereas the proposed MarkerPen is quite robust and accurate. When no reference sample is available, reference-based selection methods do not apply, but MarkerPen still shows improvement via semi-supervised marker gene selection. These findings highlight the power of MarkerPen in refining published marker genes.

3.3 Robustness

In Section 3.2 we have studied the accuracy of MarkerPen in downstream deconvolution tasks. Then a natural question is how robust MarkerPen is across different data sets. To answer this question, we experiment on the combination of four bulk data sets and three single-cell and single-nucleus reference data sets, and study the variation of their deconvolution results. Descriptions of these data sets are given in Section 2.3.

For each pair of data sets, we estimate the cell type fractions for each observation, using three marker gene selection methods: the proposed MarkerPen, the supervised method based on single-cell or single-nucleus reference data, and a fixed set of marker genes given by the **BRETIGEA** R package (McKenzie et al., 2018). Figure 4A shows the estimated fractions averaged over all observations in the data set. It is easy to see that the supervised algorithm and **BRETIGEA** generate significantly different results under three reference data sets, whereas MarkerPen is much more consistent and robust. We then compute a metric (Section S.5) to measure the variation of estimated fractions across different reference data sets, and show the values in Figure 4B. The first four panels give the comparison in each bulk data set, and the last panel shows the result over all data sets. In all settings MarkerPen is much more robust to the choice of single-cell reference data compared with others.

4 Conclusion and discussion

We have presented the MarkerPen algorithm for identifying cell-type-specific marker genes from bulk tissue data. Unlike most marker gene detection methods that heavily rely on single-cell reference samples, MarkerPen is a semi-supervised method that only requires the bulk data and a prior marker gene list. This feature makes the algorithm

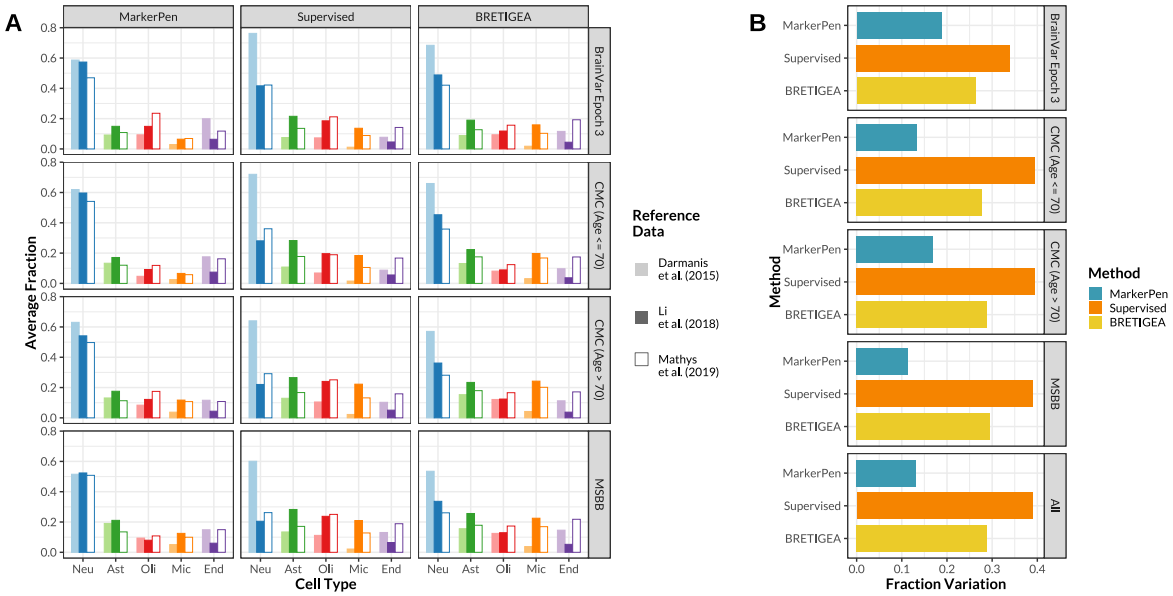

Figure 4: (A) Estimated average cell type fraction on different bulk data sets using three single-cell reference data sets and three marker gene selection methods. Deconvolution is conducted using the **dtangle** package. Neu=neurons, Ast=astrocytes, Oli=oligodendrocytes, Mic=microglia, End=endothelial. (B) Comparison of the fraction variation metric for three marker gene selection methods under various bulk data sets. This metric is used to quantify the variation of fraction estimates across different single-cell and single-nucleus reference data sets.

especially useful when tissue level data are not well matched with available single-cell data. More importantly, using well selected marker genes corrects the bias and error of downstream analyses of bulk tissue samples. Furthermore, MarkerPen interfaces nicely with other marker gene selection algorithms. For example, supervised methods applied to single-cell RNA-seq data can provide the prior gene list for MarkerPen.

A promising application of MarkerPen is to study the evolution of marker genes over developmental stages. Preliminary studies of the CMC data reveal that some marker genes identified from younger subjects are less correlated in older brains. The BrainVar data, which include brains sampled over all developmental stages, would provide an ideal data set to further investigate how marker genes change over time; however, it will be more challenging to compare marker genes of mature brains with those of fetal brains. We leave this topic for future explorations.

The use of single-cell RNA-seq has increased. However, there are drawbacks to single-cell data, including its noisy nature and the limited number of subjects from whom cell are harvested for study. By contrast, bulk transcriptome data are less noisy,

and they can readily be sampled from many subjects at a reasonable cost. With larger sample sizes, bulk tissue samples can be much more informative for downstream analyses, such as eQTL mapping. With the help of good marker genes, many deconvolution methods can provide accurate estimates of cell type fractions (Zhong et al., 2013; Gaujoux and Seoighe, 2013; Newman et al., 2015; Hunt et al., 2018; Newman et al., 2019). Furthermore, cell type fractions are input of methods such as MIND (Wang et al., 2020a) and bMIND (Wang et al., 2020b) to estimate CTS expression profiles from bulk tissue samples, permitting cell-type analysis for features such as eQTLs. The performance of these algorithms is highly dependent on the selection of good marker genes, hence MarkerPen can play a critical role in the analysis of CTS expression.

There are two limitations to the current version of MarkerPen. First, although MarkerPen is based on the eigen decomposition of correlation matrices, its computational complexity is greater than ordinary principal component analysis. In practice, one might need to limit the search range of genes to a few thousand. Despite this restriction, the algorithm has been implemented in the **markerpen** R package with core part written in efficient C++ code. Another challenge for MarkerPen is to detect cell types that are similar, such as neuron subtypes. These subtypes do not induce a strict block structure in the correlation matrix, making it harder to identify subtype-level marker genes.

MarkerPen can be extended in several directions. For instance, the current algorithm that selects marker genes performs the calculation on one cell type at a time. It may achieve better performance, however, by jointly selecting mutually exclusive marker genes for multiple cell types. Another promising direction would be to extend MarkerPen to analyzing unannotated single-cell RNA-seq data. It might be useful in selecting marker genes for clustering unlabeled cells.

Acknowledgments

We are indebted to Bernie Devlin for suggesting this topic of inquiry and Lu Xie for preliminary investigations of the idea several years ago. We benefited from comments on data analysis from Gabriel Hoffman, Michael Breen and Panos Roussos.

Data were generated as part of the CommonMind Consortium supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276,

RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881, AG02219, AG05138, MH06692, R01MH110921, R01MH109677, R01MH109897, U01MH103392, and contract HHSN271201300031C through IRP NIMH. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories, and the NIMH Human Brain Collection Core. CMC Leadership: Panos Roussos, Joseph Buxbaum, Andrew Chess, Schahram Akbarian, Vahram Haroutunian (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Enrico Domenici (University of Trento), Mette A. Peters, Solveig Sieberts (Sage Bionetworks), Thomas Lehner, Stefano Marenco, Barbara K. Lipska (NIMH).

The results published here are in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.synapse.org). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation, ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic). Additional phenotypic data can be requested at www.radc.rush.edu.

Funding

This work was supported, in part, by the National Institute of Mental Health (NIMH) grants R01MH123184 and R37MH057881. Jing Lei's research is partially supported by National Science Foundation grants DMS-1553884 and DMS-2015492.

References

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one*, 4(7):e6098.

Allen, M., Carrasquillo, M. M., Funk, C., Heavner, B. D., Zou, F., Younkin, C. S., Burgess, J. D., Chai, H.-S., Crook, J., Eddy, J. A., Li, H., Logsdon, B., Peters, M. A., Dang, K. K., Wang, X., Serie, D., Wang, C., Nguyen, T., Lincoln, S., Malphrus, K., Bisceglio, G., Li, M., Golde, T. E., Mangravite, L. M., Asmann, Y., Price, N. D., Petersen, R. C., Graff-Radford, N. R., Dickson, D. W., Younkin, S. G., and Ertekin-Taner, N. (2016). Human whole genome genotype and transcriptome data for alzheimer's and other neurodegenerative diseases. *Scientific data*, 3:160089.

Allen, M., Wang, X., Burgess, J. D., Watzlawik, J., Serie, D. J., Younkin, C. S., Nguyen, T., Malphrus, K. G., Lincoln, S., Carrasquillo, M. M., Ho, C., Chakrabarty, P., Strickland, S., Murray, M. E., Swarup, V., Geschwind, D. H., Seyfried, N. T., Dammer, E. B., Lah, J. J., Levey, A. I., Golde, T. E., Funk, C., Li, H., Price, N. D., Petersen, R. C., Graff-Radford, N. R., Younkin, S. G., Dickson, D. W., Crook, J. R., Asmann, Y. W., and Ertekin-Taner, N. (2018). Conserved brain myelination networks are altered in alzheimer's and other neurodegenerative diseases. *Alzheimer's & Dementia*, 14(3):352–366.

Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11):1969–1979.

Bakken, T. E., Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., Dalley, R. A., Royall, J. J., Lemon, T., Shapouri, S., Aiona, K., Arnold, J., Bennett, J. L., Bertagnolli, D., Bickley, K., Boe, A., Brouner, K., Butler, S., Byrnes, E., Caldejon, S., Carey, A., Cate, S., Chapin, M., Chen, J., Dee, N., Desta, T., Dolbeare, T. A., Dotson, N., Ebbert, A., Fulfs, E., Gee, G., Gilbert, T. L., Goldy, J., Gourley, L., Gregor, B., Gu, G., Hall, J., Haradon, Z., Haynor, D. R., Hejazinia, N., Hoerder-Suabedissen, A., Howard, R., Jochim, J., Kinnunen, M., Kriedberg, A., Kuan, C. L., Lau, C., Lee, C.-K., Lee, F., Luong, L., Mastan, N., May, R., Melchor, J., Mosqueda, N., Mott, E., Ngo, K., Nyhus, J., Oldre, A., Olson, E., Parente, J., Parker, P. D., Parry, S., Pendergraft, J., Potekhina, L., Reding, M., Riley, Z. L.,

Roberts, T., Rogers, B., Roll, K., Rosen, D., Sandman, D., Sarreal, M., Shapovalova, N., Shi, S., Sjoquist, N., Sodt, A. J., Townsend, R., Velasquez, L., Wagley, U., Wakeman, W. B., White, C., Bennett, C., Wu, J., Young, R., Youngstrom, B. L., Wohnoutka, P., Gibbs, R. A., Rogers, J., Hohmann, J. G., Hawrylycz, M. J., Hevner, R. F., Molnár, Z., Phillips, J. W., Dang, C., Jones, A. R., Amaral, D. G., Bernard, A., and Lein, E. S. (2016). A comprehensive transcriptional map of primate brain development. *Nature*, 535(7612):367–375.

Butler, L. M., Hallström, B. M., Fagerberg, L., Pontén, F., Uhlén, M., Renné, T., and Odeberg, J. (2016). Analysis of body-wide unfractionated tissue data to identify a core human endothelial transcriptome. *Cell systems*, 3(3):287–301.

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., Xing, Y., Lubischer, J. L., Krieg, P. A., Krupenko, S. A., Thompson, W. J., and Barres, B. A. (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *Journal of Neuroscience*, 28(1):264–278.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Gephart, M. G. H., Barres, B. A., and Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290.

De Jager, P. L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B. N., Felsky, D., Klein, H.-U., White, C. C., Peters, M. A., Lodgson, B., Nejad, P., Tang, A., Mangravite, L. M., Yu, L., Gaiteri, C., Mostafavi, S., Schneider, J. A., and Bennett, D. A. (2018). A multi-omic atlas of the human frontal cortex for aging and alzheimer's disease research. *Scientific data*, 5:180142.

Fertuzinhos, S., Li, M., Kawasawa, Y. I., Ivic, V., Franjic, D., Singh, D., Crair, M., and Šestan, N. (2014). Laminar and temporal expression dynamics of coding and noncoding rnas in the mouse neocortex. *Cell reports*, 6(5):938–950.

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., Ruderfer, D. M., Oh, E. C., Topol, A., Shah, H. R., Klei, L. L., Kramer, R., Pinto, D., Gümüş, Z. H., Cicek, A. E., Dang, K. K., Browne, A., Lu, C., Xie, L., Readhead, B., Stahl, E. A., Xiao, J., Parvizi, M., Hamamsy, T., Fullard, J. F., Wang,

Y.-C., Mahajan, M. C., Derry, J. M. J., Dudley, J. T., Hemby, S. E., Logsdon, B. A., Talbot, K., Raj, T., Bennett, D. A., De Jager, P. L., Zhu, J., Zhang, B., Sullivan, P. F., Chess, A., Purcell, S. M., Shinobu, L. A., Mangravite, L. M., Toyoshiba, H., Gur, R. E., Hahn, C.-G., Lewis, D. A., Haroutunian, V., Peters, M. A., Lipska, B. K., Buxbaum, J. D., Schadt, E. E., Hirai, K., Roeder, K., Brennand, K. J., Katsanis, N., Domenici, E., Devlin, B., and Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience*, 19(11):1442.

Gaujoux, R. and Seoighe, C. (2012). Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infection, Genetics and Evolution*, 12(5):913–921.

Gaujoux, R. and Seoighe, C. (2013). Cellmix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, 29(17):2211–2212.

Hawrylycz, M., Miller, J. A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A. L., Jegga, A. G., Aronow, B. J., Lee, C.-K., Bernard, A., Glasser, M. F., Dierker, D. L., Menche, J., Szafer, A., Collman, F., Grange, P., Berman, K. A., Mihalas, S., Yao, Z., Stewart, L., Barabási, A.-L., Schulkin, J., Phillips, J., Ng, L., Dang, C., Haynor, D. R., Jones, A., Van Essen, D. C., Koch, C., and Lein, E. (2015). Canonical genetic signatures of the adult human brain. *Nature neuroscience*, 18(12):1832.

Hickman, S. E., Kingery, N. D., Ohsumi, T. K., Borowsky, M. L., Wang, L.-c., Means, T. K., and El Khoury, J. (2013). The microglial sensome revealed by direct rna sequencing. *Nature neuroscience*, 16(12):1896.

Hilliard, A. T., Miller, J. E., Fraley, E. R., Horvath, S., and White, S. A. (2012). Molecular microcircuitry underlies functional specification in a basal ganglia circuit dedicated to vocal learning. *Neuron*, 73(3):537–552.

Hunt, G. J., Freytag, S., Bahlo, M., and Gagnon-Bartsch, J. A. (2018). dtangle: accurate and robust cell type deconvolution. *Bioinformatics*, 35(12):2093–2099.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547.

Jung, S. and Marron, J. S. (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130.

Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V., and Oldham, M. C. (2018). Variation among intact tissue samples reveals the core transcriptional features of human cns cell classes. *Nature neuroscience*, 21(9):1171.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486.

Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., Sunkin, S. M., Li, Z., Shin, Y., Zhu, Y., Sousa, A. M. M., Werling, D. M., Kitchen, R. R., Kang, H. J., Pletikos, M., Choi, J., Muchnik, S., Xu, X., Wang, D., Lorente-Galdos, B., Liu, S., Giusti-Rodríguez, P., Won, H., de Leeuw, C. A., Pardiñas, A. F., Hu, M., Jin, F., Li, Y., Owen, M. J., O'Donovan, M. C., Walters, J. T. R., Posthuma, D., Reimers, M. A., Levitt, P., Weinberger, D. R., Hyde, T. M., Kleinman, J. E., Geschwind, D. H., Hawrylycz, M. J., State, M. W., Sanders, S. J., Sullivan, P. F., Gerstein, M. B., Lein, E. S., Knowles, J. A., and Sestan, N. (2018). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, 362(6420):eaat7615.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6).

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, B. P., Bennett, D. A., Kellis, M., and Tsai, L.-H. (2019). Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337.

McKenzie, A. T., Wang, M., Hauberg, M. E., Fullard, J. F., Kozlenkov, A., Keenan, A., Hurd, Y. L., Dracheva, S., Casaccia, P., Roussos, P., and Zhang, B. (2018). Brain cell type specific gene expression and co-expression network architectures. *Scientific reports*, 8(1):1–19.

Mostafavi, S., Gaiteri, C., Sullivan, S. E., White, C. C., Tasaki, S., Xu, J., Taga, M., Klein, H.-U., Patrick, E., Komashko, V., McCabe, C., Smith, R., Bradshaw, E. M., Root, D. E., Regev, A., Yu, L., Chibnik, L. B., Schneider, J. A., Young-Pearse, T. L., Bennett, D. A., and De Jager, P. L. (2018). A molecular network of the aging human brain provides insights into the pathology and cognitive decline of alzheimer's disease. *Nature neuroscience*, 21(6):811–819.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457.

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M., and Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782.

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., and Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. *Nature neuroscience*, 11(11):1271–1282.

Ponomarev, I., Rau, V., Eger, E. I., Harris, R. A., and Fanselow, M. S. (2010). Amygdala transcriptome and cellular mechanisms underlying stress-enhanced fear learning in a rat model of posttraumatic stress disorder. *Neuropsychopharmacology*, 35(6):1402–1411.

Ryu, E. K. and Yin, W. (2017). Proximal-proximal-gradient method. *arXiv preprint arXiv:1708.06908*.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems 26*, pages 2670–2678.

Wang, J., Devlin, B., and Roeder, K. (2020a). Using multiple measurements of tissue to estimate subject-and cell-type-specific gene expression. *Bioinformatics*, 36(3):782–788.

Wang, J., Roeder, K., and Devlin, B. (2020b). Bayesian estimation of cell-type-specific gene expression per bulk sample with prior derived from single-cell data. *BioRxiv*.

Wang, M., Beckmann, N. D., Roussos, P., Wang, E., Zhou, X., Wang, Q., Ming, C., Neff, R., Ma, W., Fullard, J. F., Hauberg, M. E., Bendl, J., Peters, M. A., Logsdon, B., Wang, P., Mahajan, M., Mangravite, L. M., Dammer, E. B., Duong, D. M., Lah, J. J., Seyfried, N. T., Levey, A. I., Buxbaum, J. D., Ehrlich, M., Gandy, S., Katsel, P., Haroutunian, V., Schadt, E., and Zhang, B. (2018). The mount sinai cohort of large-scale genomic, transcriptomic and proteomic data in alzheimer's disease. *Scientific data*, 5:180185.

Werling, D. M., Pochareddy, S., Choi, J., An, J.-Y., Sheppard, B., Peng, M., Li, Z., Dastmalchi, C., Santpere, G., Sousa, A. M. M., Tebbenkamp, A. T. N., Kaur, N., Gulden, F. O., Breen, M. S., Liang, L., Gilson, M. C., Zhao, X., Dong, S., Klei, L., Cicek, A. E., Buxbaum, J. D., Adle-Biassette, H., Thomas, J.-L., Aldinger, K. A., O'Day, D. R., Glass, I. A., Zaitlen, N. A., Talkowski, M. E., Roeder, K., State, M. W., Devlin, B., Sanders, S. J., and Sestan, N. (2020). Whole-genome and rna sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. *Cell Reports*, 31(1):107489.

Xu, X., Nehorai, A., and Dougherty, J. D. (2013). Cell type-specific analysis of human brain transcriptome data to predict alterations in cellular composition. *Systems Biomedicine*, 1(3):151–160.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., and Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142.

Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M., and Liu, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC bioinformatics*, 14(1):89.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

Zou, H. and Xue, L. (2018). A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, 106(8):1311–1320.