ML4FG Final Report
Brendan Mapes (bm3019)
Iain Curtis Wagner (icw2109)

Ontology Preserving Bulk RNAseq Cell Type Deconvolution

---

1 Abstract

Here we aim to improve performance of recent cell type deconvolution methods proposed in the literature, specifically that of Wang et al. in MuSiC, by incorporating a hierarchical tree based definition of cell types that is more detailed than that which is currently utilized by the model. Currently, high level classifications of cell types are utilized in cell type proportion estimates. More specific cell types classifications, to be offered by a more detailed cell type hierarchy that is incorporated into model training, would provide more meaningful results. Researchers aiming to answer questions of disease states and disease prognoses would benefit from a model capable of more specific cell type distinction, when the data allows. Ultimately we were unable to implement an improvement upon the MuSiC model.

2 Introduction

We have investigated cell type deconvolution methods in genomics research, which aim to estimate the varying proportions of cell types present in diseased tissue samples [1]. More specifically, we aimed to develop a tree-based definition of cell types classified by the models used for this task in the current literature. This new definition of cell types and the relationships between them would allow for a more nuanced understanding of these cell type mixtures than the published research currently presents, because that research relies on high-level, general classifications of cells. The more detailed understanding of cell type mixtures offered by a tree based definition gives researchers more clarity on the larger questions that their work aims to answer, such as how particular cell type mixtures correlate to different prognoses for different patients with the same disease. Continuing the example, those researchers could better determine cancer prognosis depending on the presence of custom cells, and what immune cells get activated. Preserving information about the relationships between cell types allows for better understanding of the trajectory of cell development, contributing to a better understanding of disease states and prognoses.

To accomplish this task a linear least squares regression based approach is most often utilized. Specifically, this work is directly motivated by MuSiC, proposed by Wang et al. In this model Wang et al. weights genes showing cross-subject and cross-cell consistency [2], and maintains a basic hierarchical relationship amongst cell types. For samples with closely related cell types, MuSiC's performance showed significant improvement from previous models shown to be most effective, such as non-negative least squares regression (NNLS). MuSiC's incorporation of information on cross-subject heterogeneity in cell-type specific gene expression and information on within-cell type stochasticity of single-cell gene expression are identified by Wang et al. as the reason for this performance improvement [2].

The goal of this work was to improve upon the methods proposed by Wang et al. and the performance of MuSiC. To do so, hierarchical clustering of the single cell RNA sequencing data would be performed prior to training the model, to create a more detailed hierarchical tree of cell types and provide the classes for prediction of cell type proportions. A more detailed hierarchical tree would create a new set of categories to which cells can be classified by existing models. The level of the tree from which cell types could be predicted would be decided by the model itself, and therefore would not be dependent on any pre-selected level of the cell hierarchy. Most importantly, the more detailed predictions offered by such a model would be more meaningful to researchers using these models in practice, e.g. to determine disease prognoses.

3 Methods

Generally, single cell RNA sequence data from varying cell types is used to determine the cell type mixtures in bulk RNA sequence data for sampled tissue. In the linear model, bulk RNA sequence data was used as a linear

combination of the single cell RNA sequence data. Mathematically the model can be represented as follows, where y refers to the gene expression measure in the bulk RNA sequence, β refers to the mixture proportion, x refers to the reference gene expression, in the single cell RNA sequence data, g indexes the genes, and k indexes cell type: $y_g = \sum_k \beta_k x_{kg}$, plus an additional noise variable [2]. The inputs to the model are the single cell RNA sequence data, with an expression level for each gene. The output is the gene expression measure in the bulk RNA sequence data. The models aim to estimate β, the mixture proportion of a given cell type, in order to provide an accurate prediction of the proportion of that cell type in the tissue sample [4].

In order to incorporate a newly defined hierarchical tree of cell types, a matrix representation of the hierarchical tree must be added to the linear model. The matrix serves to regularize β, such that nodes closer in the tree would be assigned similar values of β. To create such a matrix, the hierarchical feature regression (HFR) algorithm proposed by Pfitzinger was most promising. HFR first decomposes the ordinary least squares estimator into estimates specific to each node of a hierarchical tree produced by initial clustering of input data, and then introduces a shrinkage matrix to each level of the hierarchy to produce an HFR estimator that incorporates the hierarchical relationship structure of the predictors. The shrinkage matrix shrinks coefficients for nodes closely connected on the hierarchy to a similar value [6]. A more detailed outline of the HFR strategy for regularizing the parameter β to account for hierarchical relationships between predictors can be found in Pfitzinger's paper itself, linked in the references section [6]. The corresponding R package 'HFR' would allow for incorporation of the HFR framework, for hierarchical tree-based regularization of the decomposed OLS estimator, into the cell type deconvolution methods performed in MuSiC.

To accomplish this, two types of data were necessary: 1) single cell RNA sequencing data for various cell types, x_{kg}, and 2) bulk RNA sequencing data, to provide the output gene expression measures, y_g. As outlined by Wang et al., the bulk RNA sequencing data could be constructed from the single cell RNA sequencing data. The data used for this analysis was downloaded from the MuSiC github repository (https://github.com/xuranw/MuSiC). It included single and bulk RNA sequencing data from human pancreatic islets, both healthy and diseased tissues, originally from the work of Segerstolpe et al.[7]. Prior to download this data was already prepossessed, so exploratory data analysis (EDA) was not performed. Running MuSiC and NNLS on this data for comparison was successful, and guided completely by the vignette provided by Wang et al. for the analysis from the MuSiC paper provided in the references[2].

Our initial plan for improving upon MuSiC was to implement HFR into the MuSiC model itself. The 'MuSiC' R package already provides for easy comparison of MuSiC performance to that of NNLS. Incorporating the more sophisticated methodology to incorporate information about the hierarchical relationship of the cell types as outlined by Pfitzinger, with the HFR package, seemed like a promising path towards model improvement. The difference between the 'HFR' package's hierarchical tree and that utilized in the 'MuSiC' demonstrates the difference in the level of sophistication, shown below.

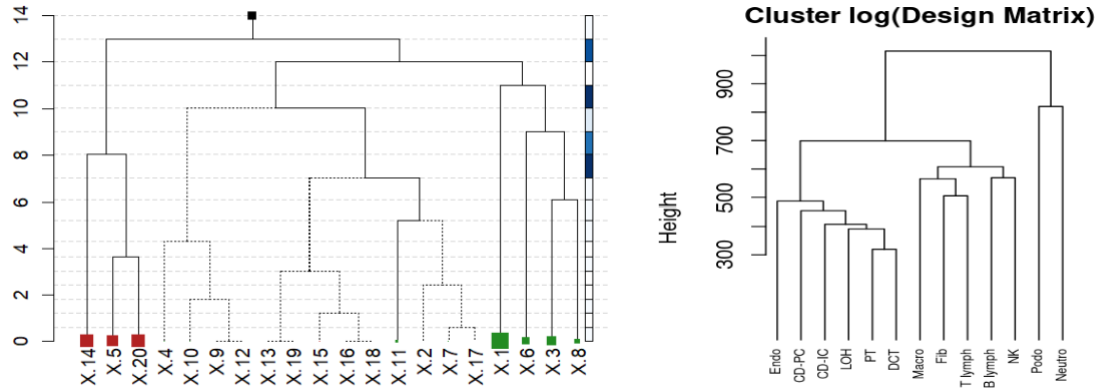

Figure 1: (Left) Example of the hierarchical dendrogram as produced by the 'HFR' R package, with leaf nodes colored and sized according to corresponding coefficients, custom edge patterns (dotted/lines) to indicate statistical significance, and a color bar on the right to indicate contribution of each level to the corresponding coefficient. (Right) Example of the hierarchical dendrogram utilized by the 'MuSiC' R package, a less sophisticated version than that provided by 'HFR'. 'MuSiC' fails to save some information on hierarchical relationships within predictors

that can be useful in informing estimations of cell type proportions.

However, we were ultimately unsuccessful getting the 'MuSiC' and 'HFR' packages to function in tandem. We still believe that incorporating Pfitzinger's strategy for regularization of the β parameter according to the hierarchical dendrogram will improve on MuSiC's performance. But we were unable to successfully implement this strategy prior to submitting this report.

4 Results

Because we ultimately failed to implement a hierarchical tree-based model for cell type deconvolution, or improve upon the performance of MuSiC, our results are limited. We did effectively run the MuSiC model on the pancreatic islets data, and compare MuSiC's performance to NNLS, by following the analysis and R code provided in the 'MuSiC' vignettes. Visual comparison of the two models' estimation of cell types are provided below, for six different cell types, and both healthy and diseased, Type 2 Diabetes (T2D), tissue samples .

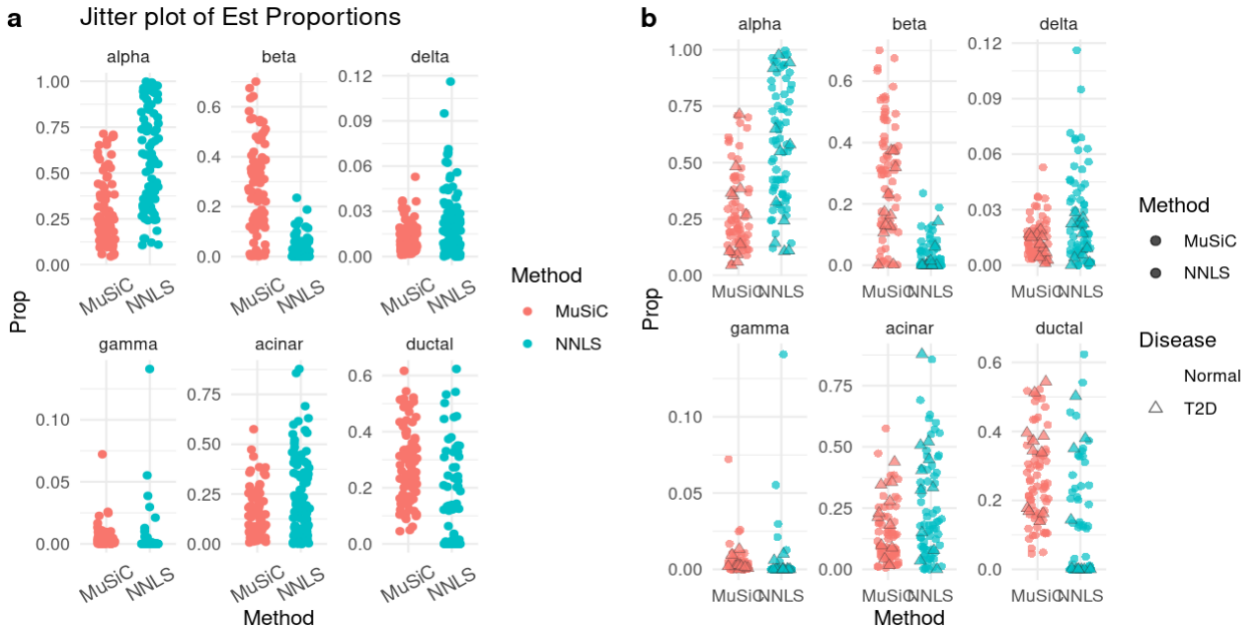

Figure 1: (a) Visual comparison of estimated cell proportions by MuSiC and NNLS models. (b) Visual comparison of MuSiC and NNLS estimated cell proportions, for normal and diseased tissue samples.

Comparison of MuSiC performance to that of NNLS on T2D samples specifically was also performed. Estimated proportions of cell types for each model were compared to ground truth cell type proportions, provided by the artificially constructed bulk data set, y_g. MuSiC achieved root mean squared error (RMSE) of 0.099, while NNLS had RMSE of 0.172. Again, we still expect some improvement in performance of MuSiC would occur, if an even more sophisticated hierarchical tree based definition of cell types were incorporated.

To further our understanding of current computational methods available for cell type convolution, and to add some breadth to this work, we performed analysis on additional data using four other models, from within the 'granulator' package from Bioconductor. Ordinary least squares (ols), support vector regression (svr), quadratic programming without constraints (qprog), quadratic programming with non-negativity and sum-to-one constraints (qprogwc), re-weighted Least Squares (rls), and a linear mixing model (dtangle) are all built in for use from within the 'granulator' package. Also built into the package are data sets required to run cell type deconvolution. As the documentation for the package states, the datasets built in to the package are "bulk RNA-seq gene expression data from peripheral blood mononuclear cells (PBMCs) from 12 healthy donors and bulk RNA-seq data of 29 isoloated immune types from 4 healthy donors,"[https://github.com/Novartis/granulator]. The subsequent analyses were guided by the vignette provided by the package, but run locally. Each type of model was run on 4 different sets of single cell RNA sequencing data to create predictions of cell type proportions in the bulk data, which was compared to the ground truth cell type proportions. NNLS and RLS show the best performance on average across the four

datasets in this analysis. We would expect to see MuSiC still outperform the seven of these models, because it maintains some information on the cell type hierarchy that these models ignore. Performance metrics for each of the models in the 'granulator' package are shown below.

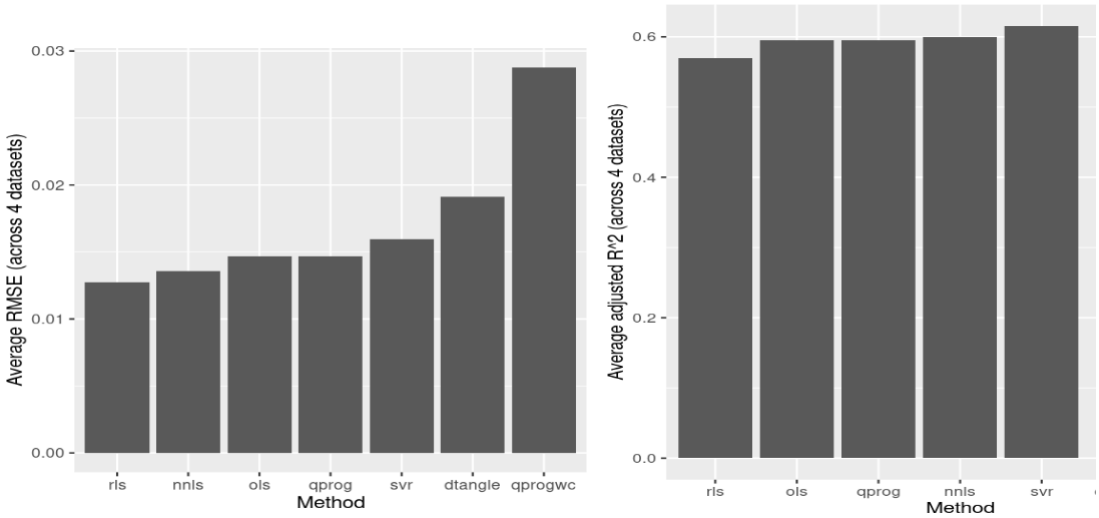

Figure 3: (Left) Average RMSE of predictions for cell type proportions of 7 model types provided by the 'granulator' R package. (Right) Average adjusted R squared for predicitons for cell type proportions of 7 model types provided by the 'granulator' R package.

Code and links to the data sets utilized for this project can be found at `https://github.com/brendanmapes/ml4fg_final`.

5 Discussion

While we failed to accomplish our original goal, incorporating a hierarchical tree based definition of cell types into a linear regression based model for cell type deconvolution, this work did develop into a high level overview of the cell type deconvolution computational methods currently available. While MuSiC appears to be most promising of the methods discussed here in this work, we still believe some improvement to MuSiC's performance would come if a more sophisticated version of hierarchical tree based definition of cell types were to be incorporated into model training. We were able to compare the peformance of MuSiC to NNLS directly on the same data, as well as compare various other models' performance on similar data for the same task of cell type deconvolution.

The choice to work from pre-built functions in the MuSiC and HFR R packages, and our inability to get these two packages working together, prevented us from accomplishing the original goal of this work. In the interim report we received feedback that our time and energy would be best spent comparing basic OLS or NNLS to a model entirely of our own building, with some hierarchical clustering algorithm incorporated there within. In that case, the creation of some matrix representation of the hierarchical tree of cell types would have needed to be of our own creation, if not implemented from the HFR algorithm. We were not confident we would be able to write and implement that algorithm to create that matrix ourselves. We did try applying the HFR strategy for creating a regularization matrix into the more basic models explored in the 'granulator' package, but were unsuccessful with that as well. The source code for the more basic models like OLS was easy enough to understand, but the source code for the HFR algorithm was not, which made incorporation of that matrix building strategy difficult. We were not confident coding this new model ourselves was feasible. Perhaps our inability to do that ourselves indicates some gap in our own knowledge that is also responsible for our inability to incorporate HFR into the MuSiC framework.

The project was significantly harder then we originally anticipated, and unfortunately outside the scope of what we could achieve this semester, but the problem still does deserve further research. In this work, we have not ruled out the approach for cluster regularization outlined by Pfitzinger as deserving of closer research by those working actively in the field of cell type deconvolution genomics research. While we were unsuccessful incorporating Pfitzinger's strategy, we still believe it may have promise.

If we were to continue this work ourselves, we would need to spend more time studying the mathematics behind Pfitzinger's approach to prepare for our own implementation of that strategy in a similar algorithm. It is also possible other pre-built functions exist in other R packages that serve a similar function, producing a matrix representation of a hierarchical tree. Once the matrix representation of the hierarchical tree of cell types is created, it can be incorporated into the linear regression based model just as would be any other regularization term. However, because MuSiC already implements a fairly sophisticated hierarchical representation of cell types into model training, it is not obvious this work would actually improve upon the performance of the MuSiC model.

6 References

[1] Li, B., Liu, J. S., amp; Liu, X. S. (2017). Revisit linear regression-based deconvolutional methods. for tumor gene expression data. Genome Biology, 18(1). doi:10.1186/s13059-017-1256-5
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1256-5

[2] Wang, X., Park, J., Susztak, K., Zhang, N. R., Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications, 10(1), 380. doi:10.1038/s41467-018-08023-
https://www.nature.com/articles/s41467-018-08023-x

[3] Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., ... Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature Methods, 12(5), 453–457. doi:10.1038/nmeth.3337
https://www.nature.com/articles/nmeth.3337#citeas

[4] Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J. C., Rodig, S., Signoretti, S., Liu, J. S., amp; Liu, X. S. (2016). Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. Genome Biology, 17(1). https://doi.org/10.1186/s13059-016-1028-7
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1028-7

[4] Pfitzinger, Johann. "Cluster Regularization via a Hierarchical Feature Regression." ArXiv.org, 7 Jan. 2022.
https://arxiv.org/abs/2107.04831

[5] Li, Bo. "Tumor Immune Estimation Resource." TIMER, http://www.cistrome.org/TIMER/download.html

[6] Pfitzinger, Johann. "Cluster Regularization via a Hierarchical Feature Regression." ArXiv.org, 7 Jan. 2022, https://arxiv.org/abs/2107.04831.

[7]Segerstolpe Å;Palasantza A;Eliasson P;Andersson EM;Andréasson AC;Sun X;Picelli S;Sabirsh A;Clausen M;Bjursell MK;Smith DM;Kasper M;Ämmälä C;Sandberg R; "Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes." Cell Metabolism, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/27667667/.