Correlation Analysis of cell-type proportion gene expression profiles estimated using RObust DEcOnvolution (RODEO)

Vaishnav Jadhav, Zahra Nozari

Functional Genomics Regensburg

Abstract

The aim of expression deconvolution is to estimate cell-type specific gene expression profiles (csGEPs) from bulk expression matrix ($Y \in M^{n_g \times n_s}$) and cell-type proportions matrix ($C \in M^{n_C \times n_S}$). Expression deconvolution is unsupervised and partial deconvolution method which only requires Y and C as input data and estimates csGEPS ($X \in M^{n_g \times n_C}$) via $Y = X \cdot C$ [2]. We found csGEPs using scRNA-seq data by RODEO algorithm and then estimated C using this csGEPs. We trained a deconvolution model with artificial training Y and C using DTD algorithm[1]. DTD is linear model which minimizes a loss function that measures the correlation between estimated C and known C. Subsequently, Pearson correlation was found out using test RNA seq-data and estimated C to determine performance of the deconvolution method. RODEO estimated X was compared with the other algorithms such as cs-LSFIT [4] and cs-QPROG [5] which utilizes least square and quadratic programming method, respectively.

Method

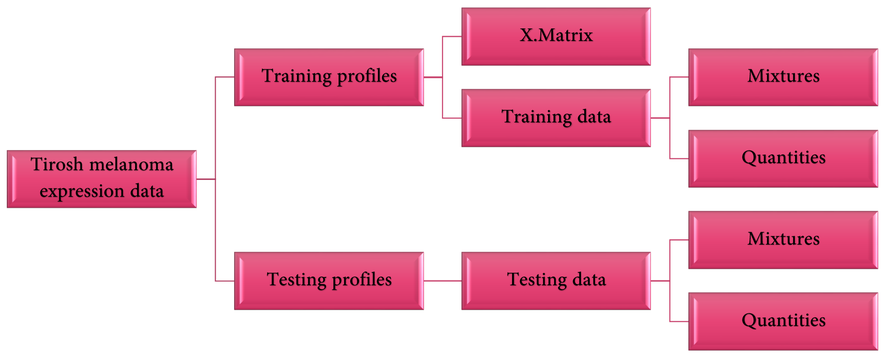
- RODEO is expression deconvolution method which evaluates for each gene across all samples by utilizing linear regression.
- RODEO utilizes regression line fitted to each gene at time by adding weight using initially all samples & cell-types.
- Robust linear model fits regression line using Huber M-estimator and objective function to be minimized is

$$\sum_{n \in N} f\left(y_{gn} - C \cdot \beta^T\right)', f(x) = \begin{cases} {}^1/_2 \cdot x^2 & |x| \le k \\ k \cdot |x| - \frac{1}{2} \cdot k^2 & |x| \le k \end{cases}$$

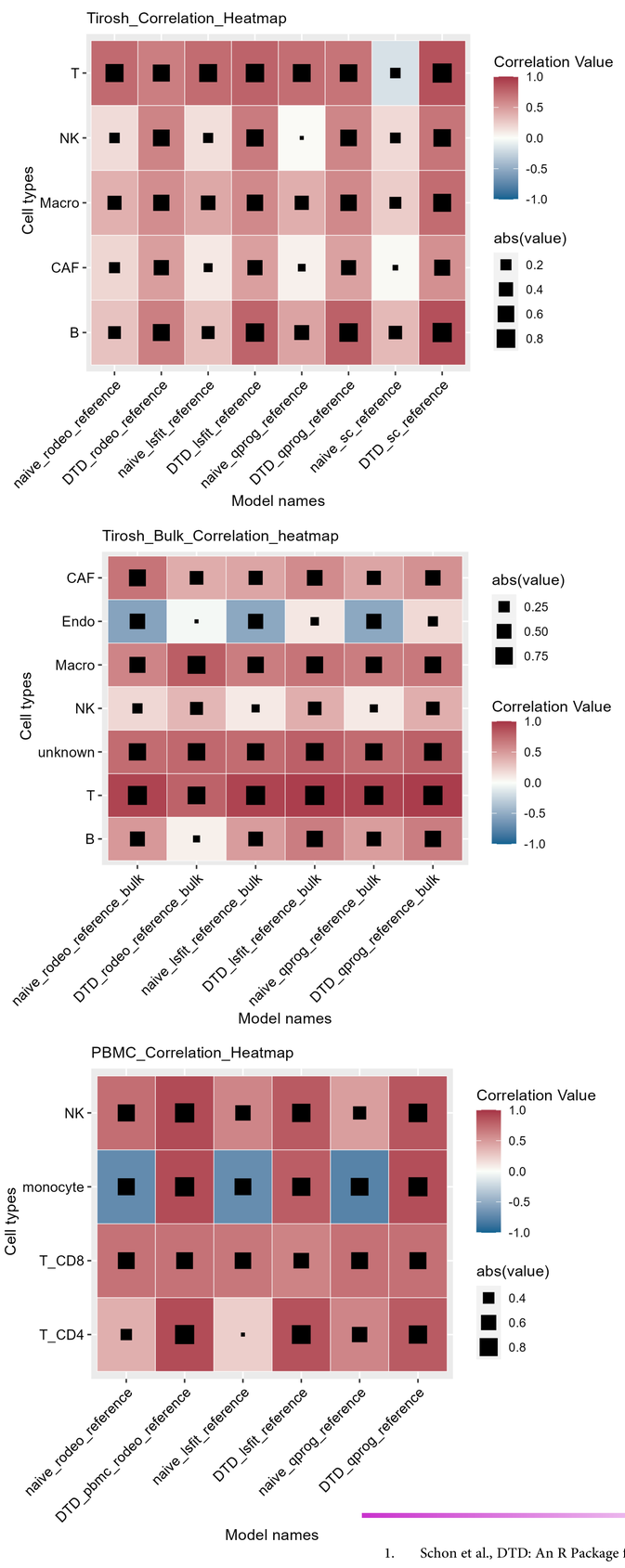
- Where, y_{gn} is measured bulk expression, β^T is unknown transposed vector to be optimized, β represents the row from matrix X telling how strongly gene expressed in each cell type.

Data

- 2 scRNA-seq data sets were used from Tirosh melanoma data and in-house PBMC data [3].
- In Tirosh data, 4645 single cell samples were isolated from 19 melanoma tumor patients and 10658 genes were analysed, whereas in PBMC data, 137 cell samples were isolated and 18563 genes.



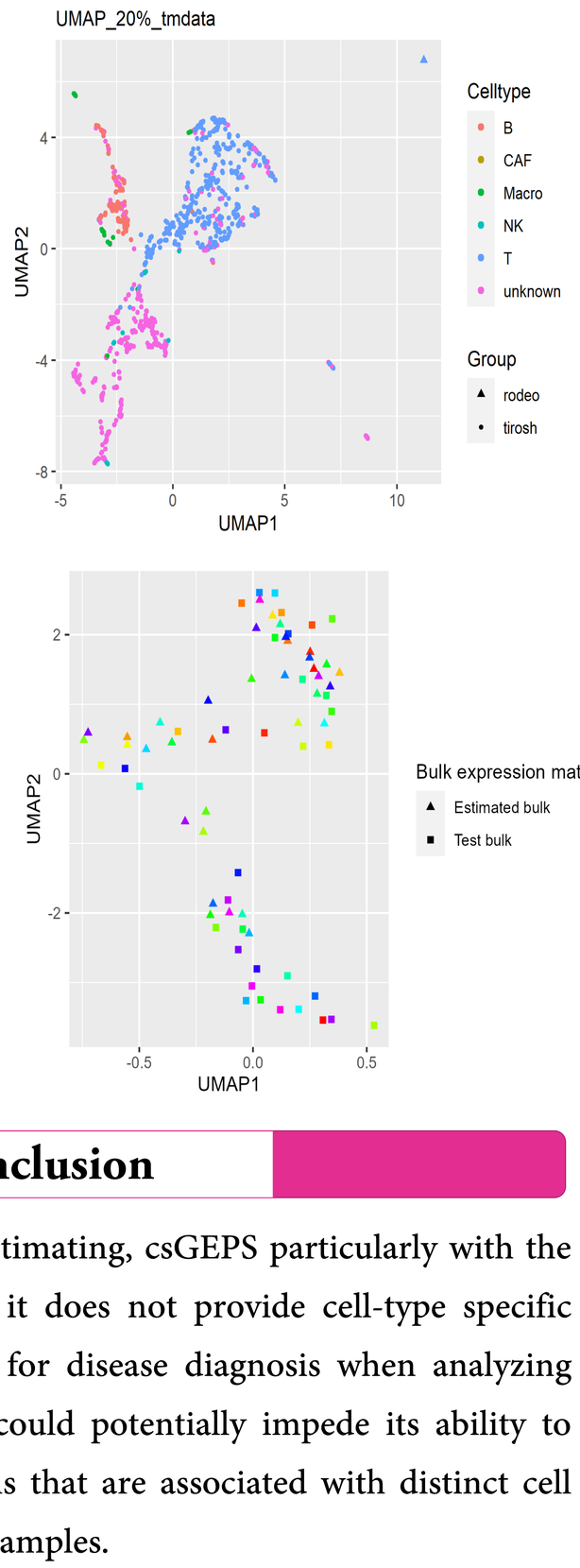
RODEO performance comparison



- Correlation between estimated C from RODEO csGEPs and test C with naïve parameter (g-vector =1) showed average correlation (~0.6-0.7) with each cell type.
- In comparison to LSFIT and QPROG estimated csGEPs, RODEO performs better with naïve parameter.
- The comparison of RODEO estimated csGEPs and Tirosh csGEPs showed that RODEO performs better than sc-data.
- Estimated cs-GEPs trained using DTD deconvolution model. With DTD csGEPs from all algorithms performs better as compare to naïve csGEPs. QPROG performs best with DTD as compared to RODEO and LSFIT.

Cell-type specific RODEO performance

- With 20% Tirosh data, it is observed that rodeo generated cell type proportions are far away from actual Tirosh cell-samples indicating that RODEO is not providing cell-type specific information across sc-samples.
- Futhermore, UMAP was performed on test Y and RODEO estimated Y. Clusters were observed of test Y and estimated Y away from each showing no cell type specific information, whereas, some test bulk and estimated bulk samples were observed close to each other



Conclusion

Although RODEO is effective in estimating, csGEPS particularly with the assistance of the DTD algorithm, it does not provide cell-type specific information that could be critical for disease diagnosis when analyzing multiple samples. This limitation could potentially impede its ability to identify specific genomic alterations that are associated with distinct cell types within heterogeneous tumor samples.

References

1. Schon et al., DTD: An R Package for Digital Tissue Deconvolution, Journal of Computational Biology, 2020
2. Elo et al, Computational deconvolution to estimate cell type-specific gene expression from bulk data, NAR Genomics, 2021.
3. Tirosh et al, Dissecting the multicellular ecosystem of metastatic melanoma by single cell RNA-seq, Cancer genomics, 2019.
4. Abbas et al., Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Erythematosus, 2009.
5. Gong et al., Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming, PLOS one , 2011.
6. Goujoux et al., CellMix: A Comprehensive Toolbox for Gene Expression Deconvolution, Bioinformatics, 2013.