Supplementary Materials for Cell type identification for single cell RNA data by bulk data reference projection

Oleg Sysoev
Department of Computer and Information Science
Linköping Univerisity
Linköping, Sweden
oleg.sysoev@liu.se

Danuta Gawel
Department of Biomedical and Clinical Sciences
Linköping Univerisity
Linköping, Sweden
danuta.gawel@liu.se

Sandra Lilja
Department of Biomedical and Clinical Sciences
Linköping Univerisity
Linköping, Sweden
sandra.lilja@liu.se

Samuel Schäfer
Department of Biomedical and Clinical Sciences
Linköping Univerisity
Linköping, Sweden
samuel.schäfer@liu.se

Mikael Benson
Department of Biomedical and Clinical Sciences
Linköping Univerisity
Linköping, Sweden
mikael.benson@liu.se

I. Bulk and single cell data relationship

To illustrate relationship between single cell and bulk data, we have done a simple experiment. Figure 1 illustrates that expression levels of single cell and bulk reference data corresponding to the same cell type have some positive correlation but the magnitude of this correlation is very low. This can be explained by both difference in technologies used for measuring single cell and bulk expression data and also great amount of noise present in the single cell data.

To justify the presence of a high amount of noise in single cell data, we have applied scVI autoencoder [1] (which makes it possible to correct for different noise characteristics of the single cell data such as varying library size or dropouts) to the single cell data used for illustration in Figure 1. The resulting relationship between the denoised expression levels obtained from scVI with and the bulk reference expressions is illustrated in Figure 2. It can be observed that the correlation between the expression profiles of the bulk data and the single cell data increases significantly if the single cell data are denoised by the scVI method.

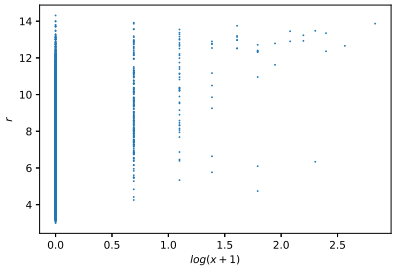

Fig. 1. Cell expressions x of a randomly selected Monocyte cell from Data Set 1 (see section III) are compared to a randomly chosen Monocyte reference vector r from the Human Cell Atlas bulk dataset (see Results section). One point at the plot correspond to one gene. The Pearson's and Spearman's correlation coefficients computed between the vectors $log(x+1)$ and r are 0.22 and 0.16 respectively.

II. RP algorithm

An algorithmic description of RP method is provided in Algorithm 1

III. Data set 1 description

10x Genomic data sets for CD4+ T cells, CD14+ monocytes, CD8+ T cells , B cells and CD56+ natural killer (NK) cells were downloaded from [2]. The downloaded single cell RNA-sequencing data was unpacked using the Matrix R package [3] and gene-cell matrices were assembled for each individual cell type. All downloaded cell type data sets provided gene expression scores for the same 32738 ensemble gene IDs. For creation of data set 1, 2500 randomly chosen cells from each cell type were combined in one gene-cell matrix. Genes in data set 1 that were not expressed in any included cell were removed, resulting in a final matrix with 17938 genes and 12500 cells.

IV. Data set 2 description

Single cell Data Set 2 was obtained by running Combine.Multiple.10X.Datasets function from **SingleR** package [4] in which we combined samples of 1000 randomly selected cells per cell subtype from 10x Genomics data set [5]. The resulting data set represents a matrix with 10000 cells and

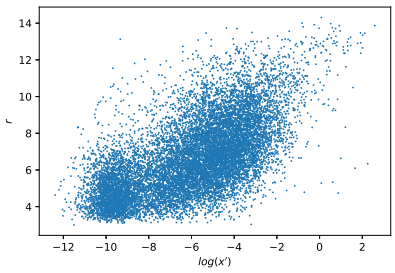

Fig. 2. Data set 1 was processed by scVI autoencoder (its settings are described in section VI). Cell expressions x of a randomly selected Monocyte cell from Data Set 1 were inputted in the fitted model and the expected mean counts x' were output by scVI for x. Expressions of $log(x')$ were compared against a randomly selected bulk Monocyte cell with expression values r. The Pearson's and Spearman's correlation coefficients computed between the vectors $log(x')$ and r are 0.63 and 0.64 respectively.

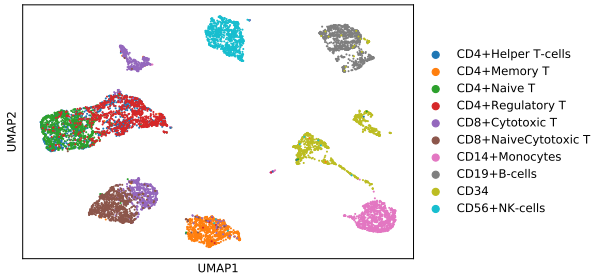

Fig. 3. UMAP representation [6] of the latent space obtained by applying the scVI model to Data Set 2. The main cell type and the alternative name of the cell type/subtype are separated by symbol "+" in the legend.

Algorithm 1 Reference Projection (RP) method

Input: Single cell data set D with n observations, reference data set R with m observations, resolution parameter s, reference data cell types $t_i, i = 1, \ldots, m$.

1. Fit scVI autoencoder to D and obtain latent space data L and denoised data D'
2. Apply Louvain clustering with resolution s to data set L and obtain K clusters.
3. Compute R' from R by applying (1)
4. Compute R'' from R' and D by applying the QQ transformation
5. Compute model (2) by using the scaled data set D'' and objective function (3)
6. For each cluster $k \in 1, \ldots, K$ assign cell type $t_{o(k)}$ matching to $r_{o(k)}$ where $o(k)$ is computed from (4).

Output: Cell type $t_{o(k)}$ for each single cell cluster $k = 1, \ldots K$

18146 genes that correspond to 10 different subtypes. Figure 3 demonstrates a visualization of these data in the latent space computed by the scVI method. It can be observed that regulatory T, naive T and helper T cells are all mixed into one cluster. This indicates that these subtypes might be hard to separate by a cell typing method. Memory T cells are, on the contrary, well separated from the remaining subtypes. CD8+ subtypes also seem to be mixed within a cluster.

V. Data set 3 description

Single cell Data Set 3 was downloaded from GEO, accession number GSE135922 [7]. This data set contains 21040 gene expressions measured for each of 4335 cells representing 10 different cell types. The cell type annotations present in these data are obtained from marker gene expressions, and we treat these annotations as ground truth cell types in our experiments (in the absence of better labels).

VI. scVI and UMAP

The scVI model in all experiments was computed by using scVI-tools Python package [8] and its function *scvi* with parameters $n\_hidden = 64$ and $n\_layers = 3$. According to our preliminary experiments, these settings led to more compact cluster representations. UMAP representations in the figures were obtained by using Python scanpy package [9] and functions sc.tl.umap with parameter $min\_dist = 0.1$ and sc.pl.umap with default settings.

s	K	CD4+	CD8+	NK-cell	Monocyte	B-cell	Overall
0.1	6	0.995	0.992	0.995	0.952	1.0	0.987
0.3	9	0.994	0.992	0.996	0.942	1.0	0.986
0.5	12	0.995	0.554	0.998	0.952	1.0	0.900
1	20	0.925	0.913	0.995	0.952	1.0	0.957
1.5	31	0.688	0.913	0.803	0.951	1.0	0.871

TABLE I
Accuracies per cell type and the overall accuracies for the RP method for various values of resolution parameter s and associated number of clusters K.

VII. Influence of resolution parameter s

We have studied the influence of the resolution parameter (and associated number of clusters K) on the accuracy of classification of the RP method when applied to Data Set 1. The resulting accuracies are reported in Table I. We observed that the accuracy of annotation for each cell type is high when the number of created clusters matches to the number of visually observed clusters, corresponding to $s = 0.1$ and $s = 0.3$. When the resolution parameter and associated amount of clusters increases some misclassifications may occur but the classification results are surprisingly stable even for the situations when a relatively big amount of clusters is produced. In particular, even when $K = 31$ clusters are created, the overall accuracy is greater than the accuracy of the alternative methods. When $K = 12$ clusters are produced, some CD8+ cells are misclassified. In fact, our results show that 32 percent of these cells are labeled as T cell gamma-delta which is still a reasonable labeling because CD8+ cells are T cells as well. Similarly, when creating $K = 31$ clusters, we have observed that 30.2% of CD4+ cells are misclassified as T cell gamma-delta.

VIII. Experiments with Data set 2

The resolution parameter was chosen as $s = 0.7$ by following the same reasoning as in the experiment with Data Set 1. Table II shows results of cell types annotations for RCA,

SingleR and RP methods. While computing this table, we computed accuracy of annotation to main cell types and also accuracy of annotation to the subtypes. If a method annotated a cell to a general cell type thus avoiding classifying to a subtype, we considered this as correct classification as well and marked the corresponding accuracies in italics. It can be observed that RCA method has low accuracy of cell type annotation for these data. In fact, we have observed that while some cell types are classified almost perfectly, many other cell subtypes are misclassified as CD8+ effector memory. SingleR and RP methods perform similarly in discovering main cell types, except of CD8+ where RP method results in higher accuracies. The same observation was done for Data Set 1. Regarding the subtype classification, RP method provides slightly higher accuracy of memory T cell annotation and also this method labels helper, naive and regulatory T cells to a more general CD4+ type while SingleR assigns more refined labels which leads though to higher misclassification rates. In addition, CD34+ cells are not classified correctly by any of the methods; instead they are assigned to either CMP or GMP cell type, both related to CD34 cell type. Overall, RP method results in higher accuracies for both main cell types and the subtypes.

IX. Experiments with data set 3

Since the bulk data had no relevant reference observations for several cell types present in the single cell data, we have performed a more detailed analysis of these data compared to the previous experiments. Instead of computing accuracies, we have computed the majority class per cell type and the percentage of the corresponding cells. More specifically, for each group of cells determined by the marker gene method, we investigate which cell types they are assigned to by a given method and what the classification proportions are per cell type. Finally, the cell type corresponding to the largest proportion was selected per each group of cells. The results are reported in Table III where the correct assignments are marked in bold. It can be observed that B cells are classified correctly by SingleR and RP methods while the accuracy is greater for RP method; endothelial cells are classified correctly by all methods while the highest accuracies are obtained by SingleR and RP methods. Macrophages are classified correctly by SingleR and RP methods, while SingleR results in an accuracy of 51% compared to 98% for RP method. Almost all Schwann cells were determined correctly again by SingleR and RP methods and T/NK class cell type determined correctly by all methods. Overall, RP method provides similar or higher annotation accuracies compared to SingleR method. Some cell types were consistently mislabeled: fibroblasts, mast cells, melanocytes, pericytes and RPE cells. This is not surprising because all of these cell types except of fibroblasts didn't have necessary references in the bulk data. Fibroblasts is a heterogeneous cell type that performs different functions in different tissues which adds to the complexity of identifying relevant references. Furthermore, recent studies found that fibroblasts are morphologically highly similar to mesenchymal stem cells in regards of immunological properties, proliferation and differentiation capacity as well as gene expression profiles [10] which may explain misclassifications of fibroblasts.

In summary, SingleR and RP methods have shown the best performance in cell type identification while RP method had notably higher accuracies for some of cell types.

References

[1] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature methods*, vol. 15, no. 12, pp. 1053–1058, 2018.

[2] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu *et al.*, "10x data," https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/, accessed: 2019-12-03.

[3] D. Bates and M. Maechler, *Matrix: Sparse and Dense Matrix Classes and Methods*, 2019, r package version 1.2-18. [Online]. Available: https://CRAN.R-project.org/package=Matrix

[4] D. Aran, A. P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R. P. Naikawadi, P. J. Wolters, A. R. Abate *et al.*, "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage," *Nature immunology*, vol. 20, no. 2, pp. 163–172, 2019.

[5] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, vol. 8, no. 1, pp. 1–12, 2017.

[6] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[7] A. P. Voigt, K. Mulfaul, N. K. Mullin, M. J. Flamme-Wiese, J. C. Giacalone, E. M. Stone, B. A. Tucker, T. E. Scheetz, and R. F. Mullins, "Single-cell transcriptomics of the human retinal pigment epithelium and choroid in health and macular degeneration. gene expression omnibus," *Proceedings of the National Academy of Sciences*, vol. 116, no. 48, pp. 24 100–24 107, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135922

[8] A. Gayoso, R. Lopez, G. Xing, P. Boyeau, K. Wu, M. Jayasuriya, E. Mehlman, M. Langevin, Y. Liu, J. Samaran, G. Misrachi, A. Nazaret, O. Clivio, C. Xu, T. Ashuach, M. Lotfollahi, V. Svensson, E. da Veiga Beltrame, C. Talavera-Lopez, L. Pachter, F. J. Theis, A. Streets, M. I. Jordan, J. Regier, and N. Yosef, "scvi-tools: a library for deep probabilistic analysis of single-cell omics data," *bioRxiv*, 2021. [Online]. Available: https://www.biorxiv.org/content/early/2021/04/29/2021.04.28.441833

[9] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, no. 1, pp. 1–5, 2018.

[10] M. Soundararajan and S. Kannan, "Fibroblasts and mesenchymal stem cells: Two sides of the same coin?" *Journal of cellular physiology*, vol. 233, no. 12, pp. 9099–9109, 2018.

Method	Monocytes	B	CD34+	Helper T	Memory T	Naive T	Regulatory T	NK	Cytotoxic T	Naive Cytotoxic T	overall
RCA main	0.98	0.99	0.0	0.0	0.0	0.0	0.0	0.0	0.99	1.0	0.40
SingleR main	0.96	0.99	0.0	0.98	0.98	0.98	0.97	0.97	0.37	0.25	0.74
RP main	0.95	0.99	0.0	0.98	0.96	0.98	0.98	0.99	0.57	0.98	0.84
RCA sub	0.98	0.99	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.20
SingleR sub	0.96	0.99	0.0	0.0	0.90	0.68	0.0	0.97	0.0	0.0	0.45
RP sub	0.95	0.99	0.0	*0.98*	0.96	*0.98*	*0.98*	0.99	0.0	0.0	0.68

TABLE II

ACCURACIES PER SUBTYPE AND THE OVERALL ACCURACIES OBTAINED BY RCA, SINGLER AND RP METHODS FOR DATA SET 2

Cell Type	RCA ML	RCA	SingleR ML	SingleR	RP ML	RP
B	T	1.0	**B**	0.89	**B**	0.98
Endothelial	**Endothelial**	0.88	**Endothelial**	0.96	**Endothelial**	0.98
Fibroblast	Tissue Stem	0.99	Tissue Stem	0.70	Tissue stem	0.60
Macrophage	Monocyte	0.91	**Macrophage**	0.51	**Macrophage**	0.98
Mast	T	1.0	NK	0.47	B	1.0
Melanocyte	Tissue Stem	0.67	Schwann	0.52	Embryonic stem	0.51
Pericyte	Tissue Stem	0.99	Tissue Stem Cells	0.96	Tissue Stem	0.92
RPE	Tissue Stem	0.86	Astrocytes	0.28	Neurons adrenal medulla	0.98
Schwann	Tissue Stem	0.96	**Schwann**	0.94	**Schwann**	0.94
T/NK	**T**	1.0	**T**	0.60	**T**	1.0

TABLE III

MAJORITY LABEL (ML) AND THE PROPORTION OF CELLS ASSIGNED TO THAT LABEL PER CELL TYPE OBTAINED BY RCA, SINGLER AND RP METHODS FOR DATA SET 3