Optimized semi-deconvolution using reference data

Alan Luo

Computational Biology Department, School of Computer Science, Carnegie Mellon University

SCS Honors Thesis

Dr. Russell Schwartz

April 30, 2024

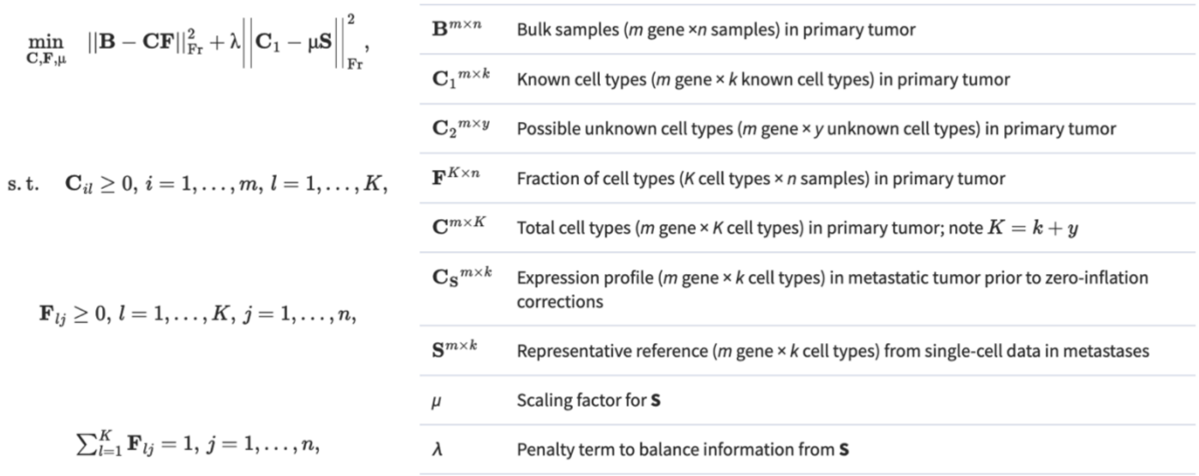
Abstract

Robust and Accurate Deconvolution Single-Cell, or RADs, is an algorithm for integrating bulk and single-cell genomic data in cancer progression studies. Methods like RADs are used to examine the composition of cells making up a tumor and how their behavior is perturbed in different tumor sites, which in turn yields insight to how cancers might be better monitored or treated. RADs is an algorithm that performs a technique called semi-deconvolution, which seeks to infer frequencies of cell types and their gene expression evolution over stages of cancer progression. The first efforts in this research area focused on specific use cases wherein the data that one has available is limited to bulk data profiling average genomic features of mixtures of many distinct cells. Single-cell data has revolutionized the field by allowing one to track genomic behavior of individual cells in a tumor but is not always technically feasible. There are situations when one has samples suitable for single-cell methods, such as some recent metastases, but also samples only suitable for bulk methods, such as biopsies of archived primary tumors that may have been preserved years earlier. RADs focuses on such scenarios but can have poor resolution for identifying and quantifying the many different cell types that may be found in the bulk data. Hence, the goal of this study was to improve upon RADs by making use of reference single-cell RNA-seq datasets that provide models of gene expression of many known cell types. At present, several new combinations of data have been explored to improve the algorithm, using different penalty weights each time. The results were that the performance worsened as the penalty weight increased, at least for one of the combinations. In addition, the changes in cell type compositions observed across the penalty weights were somewhat consistent with expectations from prior biological knowledge. The results indicate that the prior reference-free RADs method could be adapted to accommodate third-party reference data sources. More results are being collected.

Introduction to the problem

The goal of this study was to optimize the semi-deconvolution algorithm Robust and Accurate Deconvolution Single-Cell, or RADs. More specifically, this optimization would maximize the ability of RADs to use external reference data to interpret bulk genomic RNA data as mixtures of cell types and infer the clonal fractions describing the compositions of these cell types and gene expressions describing activity of gene networks active in each cell type. This optimization would reveal more about how the composition and activity of tumors change across stages of progression, most notably the transition from primary tumors to metastases. This would help to reveal the mechanisms causing certain tumors to be more aggressive, thereby providing greater benefit to cancer researchers in developing treatments.

The precise mathematical problem being solved by RADs is the constraint optimization problem shown in Fig. 01. This information was drawn directly from a previous study on RADs which introduced this algorithm (Lei et al., 2022).



$$\min_{\mathbf{C},\mathbf{F},\mu} \quad ||\mathbf{B}-\mathbf{CF}||_{\mathrm{Fr}}^2 + \lambda \left|\left|\mathbf{C}_1 - \mu\mathbf{S}\right|\right|_{\mathrm{Fr}}^2,$$

$$\text{s.t.} \quad \mathbf{C}_{il} \geq 0,\, i = 1,\ldots,m,\, l = 1,\ldots,K,$$

$$\mathbf{F}_{lj} \geq 0,\, l = 1,\ldots,K,\, j = 1,\ldots,n,$$

$$\textstyle\sum_{l=1}^{K} \mathbf{F}_{lj} = 1,\, j = 1,\ldots,n,$$

$\mathbf{B}^{m \times n}$	Bulk samples (*m* gene ×*n* samples) in primary tumor
${\mathbf{C}_1}^{m \times k}$	Known cell types (*m* gene × *k* known cell types) in primary tumor
${\mathbf{C}_2}^{m \times y}$	Possible unknown cell types (*m* gene × *y* unknown cell types) in primary tumor
$\mathbf{F}^{K \times n}$	Fraction of cell types (*K* cell types × *n* samples) in primary tumor
$\mathbf{C}^{m \times K}$	Total cell types (*m* gene × *K* cell types) in primary tumor; note $K = k + y$
${\mathbf{C}_\mathbf{S}}^{m \times k}$	Expression profile (*m* gene × *k* cell types) in metastatic tumor prior to zero-inflation corrections
$\mathbf{S}^{m \times k}$	Representative reference (*m* gene × *k* cell types) from single-cell data in metastases
μ	Scaling factor for **S**
λ	Penalty term to balance information from **S**

Fig. 01. Left scheme depicts the optimization problem depicted in its constraint optimization problem form. Right scheme lists the meanings behind each of the terms in the optimization problem. Specifically, the B, S, C, and F terms are the primary terms with the B term signifying the bulk dataset, the S term signifying the reference single-cell dataset, the C term signifying the inferred gene expression profile, and the F term signifying the inferred clonal fraction profile. The C term is further divided based upon whether the cell types are known or not. In addition, there are two regularization parameters which are the scaling factor on the reference data and the penalty weight on the difference between the reference data and the inferred gene expression profile, signified by μ and λ respectively. Finally, the imposed constraints allow for the inferred C and F matrices to represent what they are intended to represent as they are required to be nonzero and, in the case of F, sum to 1. In addition, the constraints allow them to describe the bulk data as evidenced by their dimensions.

Background

Semi-deconvolution is a strategy for inferring how cell type compositions and their behavior change between different tissue samples, such as between a primary tumor and one or more metastases in the same patient. Typically, the concern is in how clonal fractions and gene expressions evolve over stages of tumor progression. The term semi-deconvolution comes from the hybrid use of both bulk RNA-seq data and single-cell RNA-seq data, in contrast to purely deconvolution approaches in which one seeks to infer single-cell behavior from only mixed, or bulk, data.

Using both bulk and single-cell data rather than solely the bulk data is expected to lead to greater accuracy to the deconvolution, on the assumption that some portion of the mixed data can be explained by observed single cells. In addition, although single-cell approaches pose several advantages over purely bulk data approaches for their high accuracy and low costs (Kuipers & Beerenwinkel, 2017; Lim et al., 2020), single-cell approaches still suffer from several pitfalls as single-cell data remains inaccessible for older samples and lacks the level of comprehensive databases seen with bulk data such as International Cancer Genome Consortium (Naxerova & Jain, 2015; Zhang et al., 2011). Hence, there is a need for a more hybrid approach incorporating both bulk and single-cell data.

The RADs algorithm was limited for performing only on data that comes from the same patient (Lei et al., 2022). Although the reference and bulk data each come from different sites throughout the human body, the requirement is that the data ultimately comes from the same patient. However, there are now many studies of single-cell data in healthy tissues and datasets of such reference data on research subjects that might be used to help interpret data on other subjects. These reference libraries provide far more examples of cells with known cell types and other annotations, enabling for more comprehensive, if somewhat biased, models of these known cell types in a new subject. By having the reference data come from these libraries, it is hoped that the algorithm will be more effective in identifying the components of tumors that are well explained by cells of known type, better enabling the analysis of their composition and the resolution of the patient-specific portions of tumors that largely correspond to novel cell types evolving within the tumor, also known as tumor clones.

Methodology

At a high level, the work consisted of evaluating the performance of RADs on different combinations of bulk, single-cell, and reference data and then modifying the algorithm as needed for each combination. The combinations that were explored are listed in the results section and are herein referred to by a unique numerical identifier beginning from 01. For example, Combination ID01 refers to the data corresponding to the ID of 01. During pre-processing of the data, only the genes that overlap between the bulk and reference datasets were considered, as the optimization problem requires common dimensions for the matrices.

The result produced from RADs is an inferred clonal fraction matrix named F, which describes the fraction of each cell type inferred to be present in each bulk sample, and an inferred gene expression profile matrix named C, which describes the activity level of each gene in each cell type. These two matrices describe the projected fraction evolution and the expression evolution of the bulk dataset, respectively.

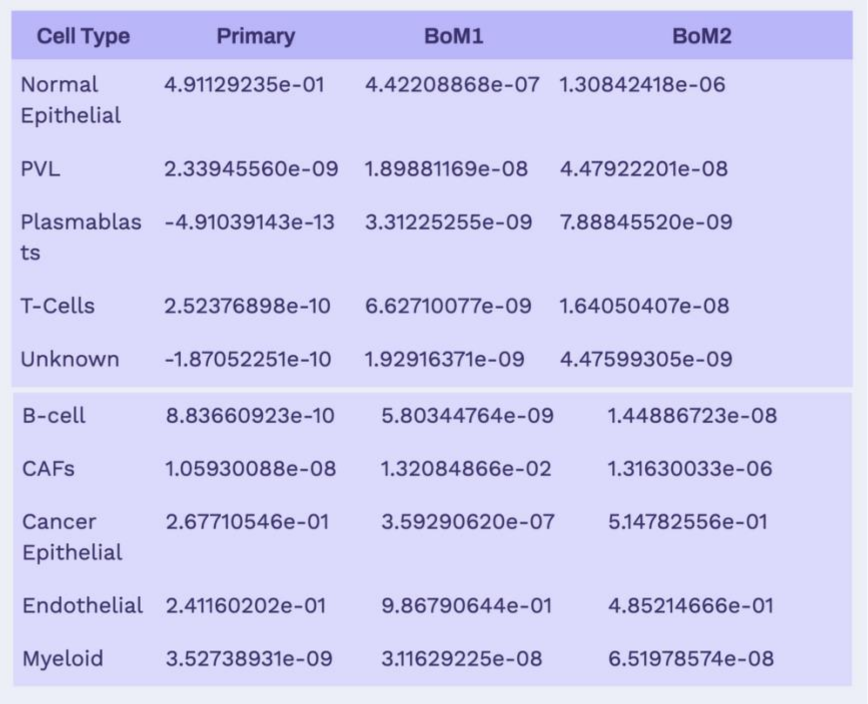
Once these results were collected, three methods were then employed to validate the results. **One validation method** was to visualize the expression profile using heatmaps which provides a way of assessing consistency between inferred expression values and single cell data from the reference or the same patient. The heatmap was then examined to determine the existence of any trends in expression, such as increasing in expression across two cells followed by decreasing across the next cell then returning to the same level across the next cells. The existence of a pattern is important because for any gene, the exact way it is expressed can vary depending upon the cell. Regardless of the way it is expressed, however, the effect should be consistent for similar cells, such as T-cells or cancer epithelial cells. In addition, the genes that were associated with high levels of expression were examined to determine if they played important roles in cancer, such as in cell cycle regulation. Hence, the heatmaps were analyzed to determine any possible clones or groups of related cells which would then be compared to the known reference data to check if the inferred clones matched the actual cell types. **Another validation method** was to analyze the changes in clonal fractions to determine if they could also be justified with a biological explanation. For example, if there was a large observed increase in clonal fractions from the primary site to one of the metastasis sites, then the cell type corresponding to that increase should require such an increase, as dictated by the pathways and biological processes to which that cell type contributes or the tissue type in which the metastasis occurs.

In the deconvolution problem, there are two regularization terms. One of the terms is a penalty weight associated with the difference between the computed reference matrix S and the given single-cell matrix C, which is used to bias the inference towards explaining the bulk data with cell types previously observed in single-cell data. The other term is a scaling factor of the reference data S, which needs to be learned due to the different technologies involved in making bulk versus single-cell gene expression measurements. Varying weights for the penalty weight were used so that the effects of the penalty weight on the resulting inference could be examined. From there, modifications to the algorithm can then be made, such as rearranging the penalty weight to another term within the optimization expression. The scaling factor is optimized during the RADs algorithm and is thus left alone unless the way it was optimized ever required changing.

Results

ID	Bulk RNA-seq Dataset	Single-Cell RNA-seq Reference Dataset
01	Single cell RNA analysis of breast cancer bone metastases (GEO: GSE190772).	A single-cell and spatially resolved atlas of human breast cancers (Broad Institute Single-Cell Portal).
02	Simulated dataset from ID01 Reference Dataset composed of 415 randomly chosen genes and 10 randomly chosen cells. The cells are of the following types, in order: T-cell, cancer epithelial, plasmablast, CAF, myeloid, plasmablast, CAF, T-cell, cancer epithelial, T-cell.	ID01 Reference Dataset using only the genes from ID02 Bulk Dataset.

Fig. 02. Table displaying each of the datasets examined. Each pair of datasets is identified using a numerical identifier beginning from 01. For each pair, a bulk dataset and a single-cell reference dataset was used.



Cell Type	Primary	BoM1	BoM2
Normal Epithelial	4.91129235e-01	4.42208868e-07	1.30842418e-06
PVL	2.33945560e-09	1.89881169e-08	4.47922201e-08
Plasmablasts	-4.91039143e-13	3.31225255e-09	7.88845520e-09
T-Cells	2.52376898e-10	6.62710077e-09	1.64050407e-08
Unknown	-1.87052251e-10	1.92916371e-09	4.47599305e-09
B-cell	8.83660923e-10	5.80344764e-09	1.44886723e-08
CAFs	1.05930088e-08	1.32084866e-02	1.31630033e-06
Cancer Epithelial	2.67710546e-01	3.59290620e-07	5.14782556e-01
Endothelial	2.41160202e-01	9.86790644e-01	4.85214666e-01
Myeloid	3.52738931e-09	3.11629225e-08	6.51978574e-08

Fig. 03. Table showing the cell types present in ID01 penalty 0 and the associated changes in clonal fractions. The columns are the three samples present within the bulk dataset: primary is the primary site, BoM1 is a bone metastasis site, and BoM2 is another bone metastasis site.

The data explored in this study are all listed in Fig. 02. Different combinations of data were used to provide better training to the RADs algorithm. Below, the results for each combination are broken down in order of the inferred clonal fractions then the inferred gene expression profile. For the fraction profile, the fractions are provided followed by an evaluation of the numbers to determine how biologically feasible they seem. For the gene expression profile, the expressions are provided followed by heatmaps and plots depicting the variations amongst the different regularization parameters, namely the penalty weight and the scaling factor.

ID01 results

In ID01 using a penalty weight of 0, most of the clonal fraction changes (Fig. 03) appear reasonable given prior biological knowledge. The trend for Normal Epithelial cells, which shows a substantial drop between primary tumor and metastasis, is explained by the fact that the metastasis is in non-epithelial tissue, namely bone marrow tissue. The trend for PVL, or Perivascular-Like Cells, showing an increase from primary to metastatic tumors, is explained by the fact that PVL are associated with angiogenesis which is blood vessel growth, since blood vessels line the bone marrow (Wu et al., 2020). The trend for Plasmablasts, also showing notable increase from primary to metastasis, is explained by the fact that plasmablasts migrate to the bone marrow once produced within the thymus, causing an increase in clonal frequency (Chu & Berek, 2013). The trend for B-cells, again showing increases in metastases relative to primary, is explained by the fact that many B-cells reside in the bone marrow regardless of whether an immune response is needed or not (Agrawal et al., 2013). The trend for CAF, or Cancer-Associated Fibroblast, which also increase in metastases although by very different degrees in the two, might be explained by the fact that CAF promotes tumor progression. Initially, CAF represses progression by forming gap junctions between activated fibroblasts (Cirri & Chiarugi, 2011). Overall, however, the rate at which CAF promotes progression exceeds that at which it represses progression. Hence, the levels of CAF might be expected to increase in the metastasis sites since the cancer has progressed at that point. The increasing trend for myeloid in metastases versus primary is also explained by the fact that myeloid is produced in the bone marrow (Galán-Díez, Cuesta-Domínguez, & Kousteni, 2018).

ID02 results

In ID02, the inferred clonal fraction profile remained constant across all the penalty weights (Fig. 04). There seemed to be no biological reason for the trend, as the peaks and dips fail to align with a consistent biological pattern. For example, although Cells 2 and 4 are plasmablasts, Cell 8 is a T-cell, raising the question of why Cells 0 and 6 failed to reach as high of a peak since they are both also T-cells.

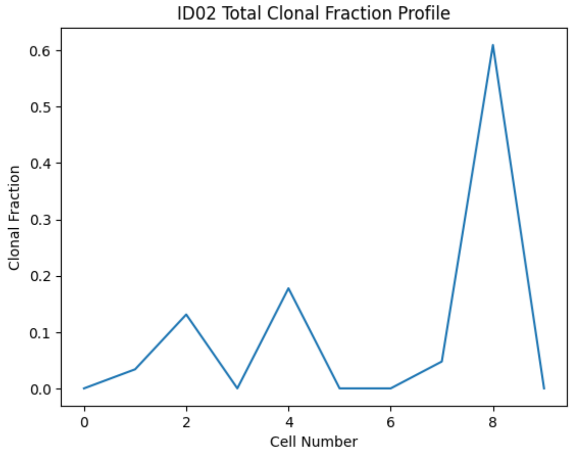


Fig. 04. The clonal fraction profile for each penalty weight. The profile was identical across all penalty weights. The cell number indicates a unique cell from the ten randomly chosen cells for the simulated bulk dataset.

In ID02 using a penalty weight of 0, the inferred gene expressions were somewhat biologically feasible as evidenced using a heatmap (Fig. 05). The genes showing high expression within the randomly selected cells were often related to cell division, a process closely associated with cancer: PSME1, SRSF5, GPX1, DOCK7, JUN, TNFAIP3, and TMEM165 (Martin et al.). However, there was not a clear pattern in gene expression levels across cells as visualized within the two heatmaps. There were promising signs in both heatmaps, namely Genes 14 and 19 in the left heatmap and Genes 3 and 25 in the right heatmap. The cells corresponding to the changes in expression, however, lacked a consistent pattern, as in the left heatmap for example, the cells that were plasmablasts corresponded to the highest expression (white color) but also a much lower expression (orange color). The aforementioned significant genes also appeared to have no particularly strong relationship with the cells. In the left heatmap for example, Genes 14 and 19 corresponded to PSME1 and SRSF5. PSME1 is a proteasome while SRSF5 is a splicing factor, neither of which bears a particularly strong association with the cells. Both these proteins have associations with fundamental biochemical processes, but neither of these processes suggests the strength of association implied by the high degree of gene expression displayed for these two genes. Hence, clones for the cells could not be determined with enough accuracy.

For the remaining penalty weights for ID02, the results were similar but with subtly different nuances in the gene expression patterns. There were still no significant gene expression patterns, meaning that identifying clones still could not be determined with sufficient accuracy. However, most of the highly significant genes themselves exhibited strong associations with biochemical processes pivotal to tumor aggression, suggesting that the reference data still provided at least some benefit to the RADs algorithm.

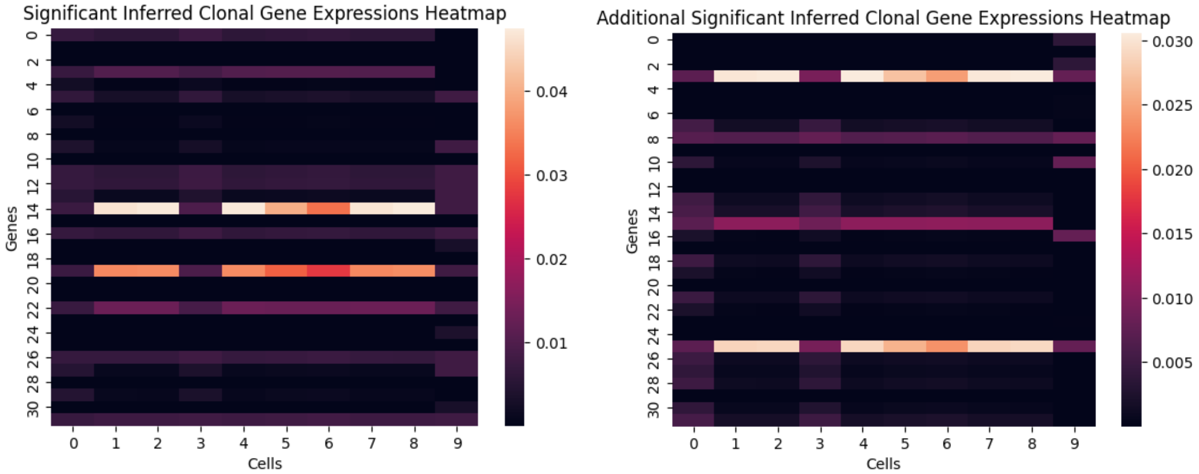


Fig. 05. Two heatmaps depicting the two regions of the overall heatmap for ID02 penalty 0 that exhibit significant gene expression as evidenced by the lighter colors. The gene numerical identifiers lining the rows are numbered relative to the region, not the overall heatmap. The cell numbers correspond to the same 10 randomly chosen cells.

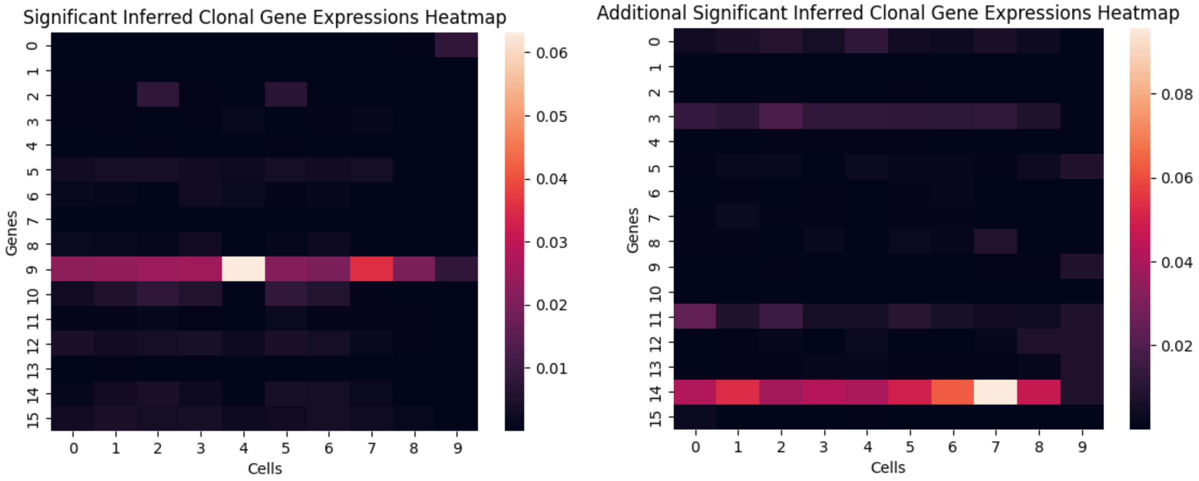


Fig. 06. Two heatmaps depicting the two regions of the overall heatmap for ID02 penalty 0.01 exhibiting significant gene expression as evidenced by the lighter colors.

In ID02 penalty 0.01, the gene expressions also seemed to be only somewhat feasible as evidenced using heatmaps (Fig. 06). There appeared to be no significant pattern in gene expression levels. However, the most strongly expressed genes, namely Gene 9 in the leftmost heatmap and Gene 14 in the rightmost heatmap of Fig. 05, were associated with cell membrane regulation and immunity: ARFRP1 and MCOLN2 (Martin et al., The Uniprot Consortium).

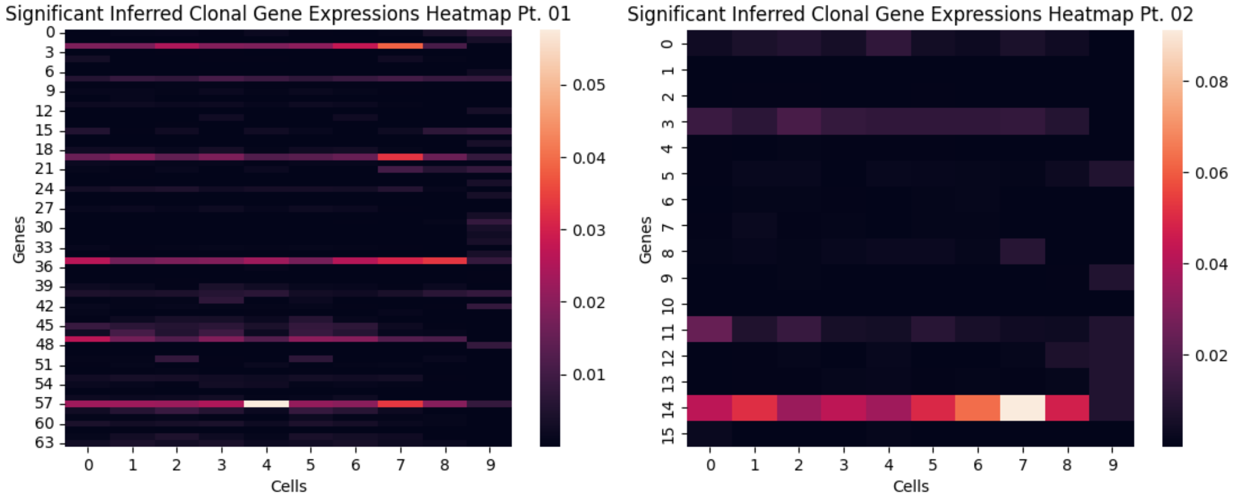


Fig. 07. Two heatmaps depicting the two regions of the overall heatmap for ID02 penalty 0.1 exhibiting significant gene expression as evidenced by the lighter colors.

For ID02 penalty 0.1, there appeared to be no significant gene expression trends (Fig. 07). However, the most strongly expressed genes as indicated by the light-colored bars in the two heatmaps, were often highly associated with fundamental components of the cell such as the transmembrane or with basic biochemical pathways such as respiration and translocation, making these genes important to processes leading to cancer: COX7A2, SEC62, PSME1, SRSF5, GPX1, TMEM210 (Sayers et al.).

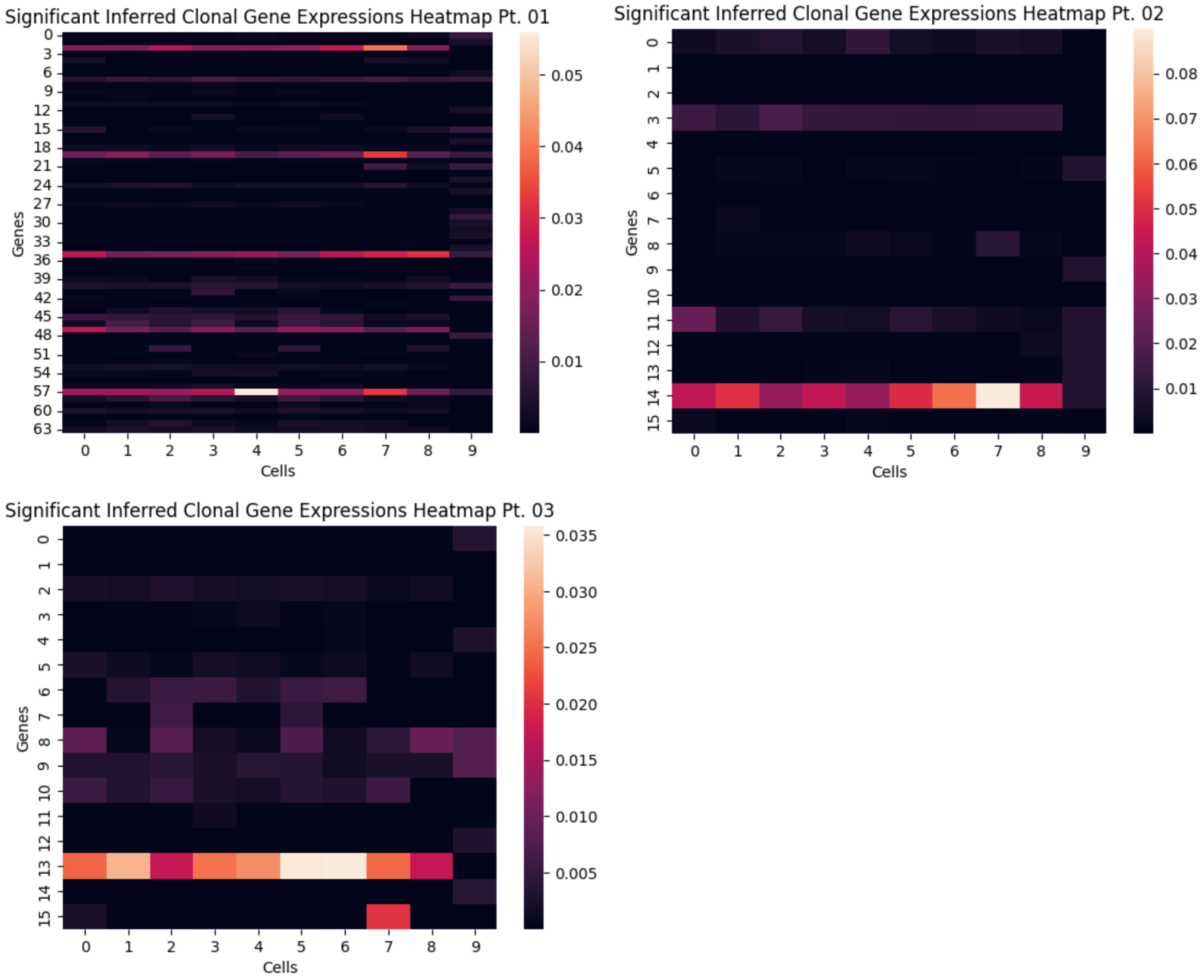


Fig. 08. Three heatmaps depicting the three regions of the overall heatmap for ID02 penalty 1 exhibiting significant gene expression as evidenced by the lighter colors.

For ID02 penalty 1, there seemed to be no particularly noticeable gene expression trend as well (Fig. 08). However, most of the highly significant genes were also significant for prior penalty weights and were associated with essential biochemical processes such as the processing of histones and transportation of amino acids. Furthermore, several genes were protooncogenes (JUN and FOSB). The significant genes were as follows: COX7A2, SEC62, PSME1, SRSF5, GPX1, SNRPG, JUN, SLC38A1, FOSB, IGHV4-59 (Sayers et al.).

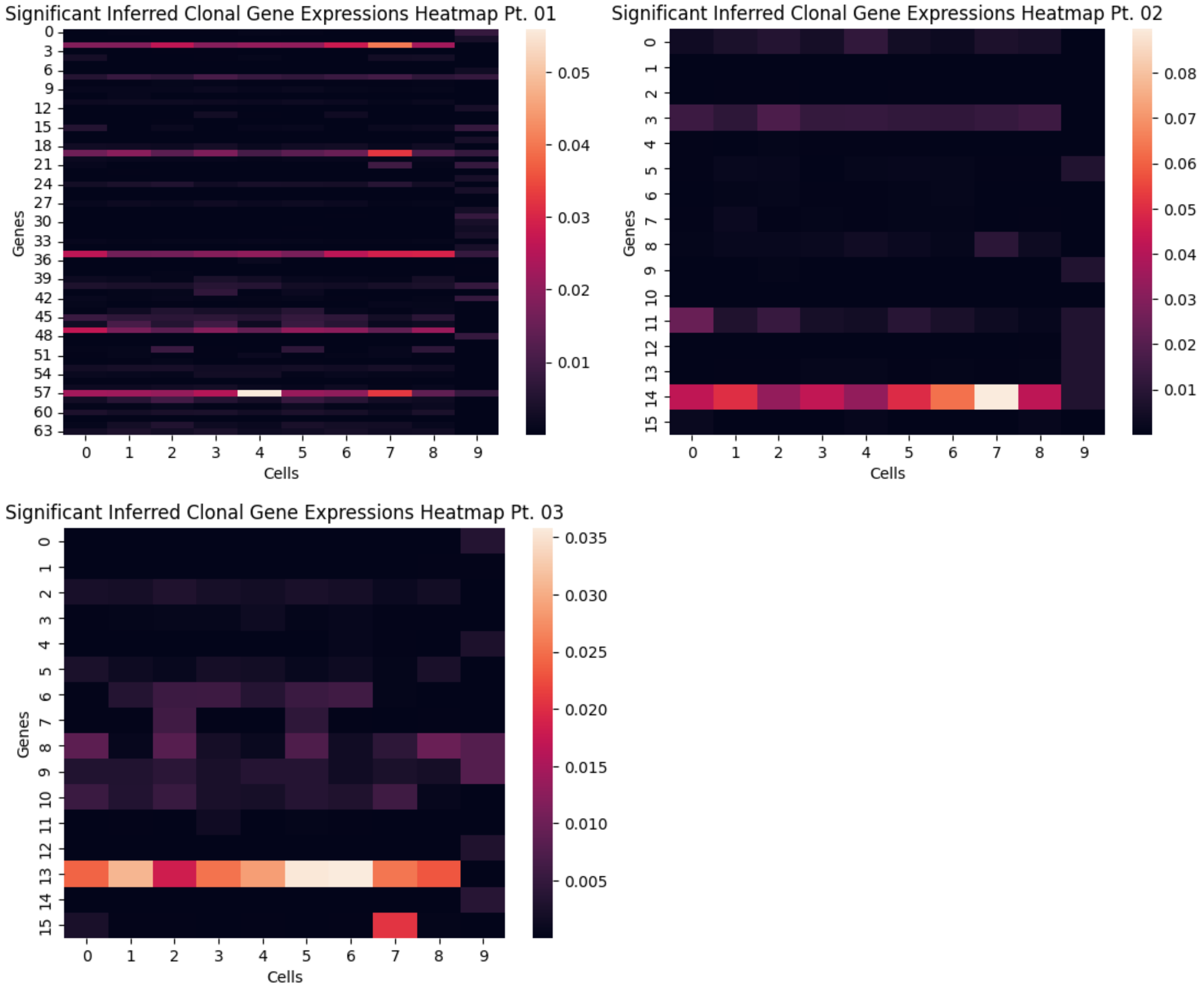


Fig. 09. Three heatmaps depicting the three regions of the overall heatmap for ID02 penalty 10 exhibiting significant gene expression as evidenced by the lighter colors.

For ID02 penalty 10, there seemed to be no significant gene expression trends (Fig. 09). However, the highly significant genes were also significant for prior penalty weights and were involved with important biochemical processes, such as membrane transport, as well as gene regulation which is associated with cancer: COX7A2, SEC62, PSME1, SRSF5, GPX1, SNRPG, SMIM14, JUN, OSBPL1A, SLC38A1, FOSB, IGHV4-59.

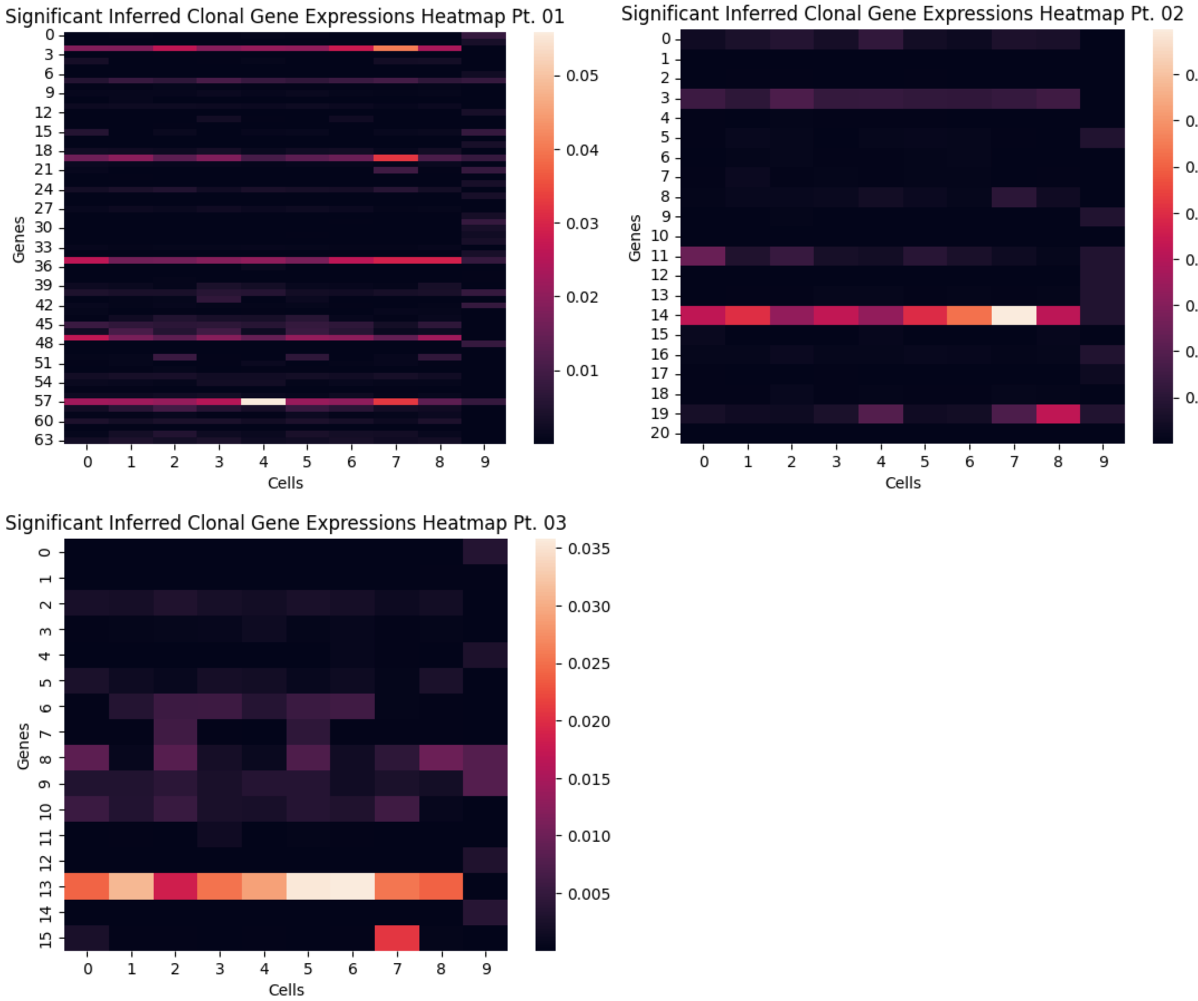


Fig. 10. Three heatmaps depicting the three regions of the overall heatmap for ID02 penalty 100 exhibiting significant gene expression as evidenced by the lighter colors.

For ID02 penalty 100, there seemed to be no significant gene expression trends (Fig. 10). However, the highly significant genes were also significant for prior penalty weights and were involved with similarly important biochemical processes: COX7A2, SEC62, PSME1, SRSF5, GPX1, SNRPG, SMIM14, JUN, TNFAIP3, OSBPL1A, SLC38A1, FOSB, IGHV4-59.

For ID02 penalty 1000, the heatmaps were virtually identical to penalty 100 and thus yielded the same lack of gene expression patterns and significant genes.

As evidenced by the gene expression profiles, the penalty weight seemed to have somewhat of an impact at best on the inferred gene expressions. The most significant differences were from 0 to 0.01 and from 0.01 to 0.1 while the differences in greater penalties were non-significant. The difference from 0 to 0.01 was notable for the loss of a visual pattern in the graph as there was a pattern of increasing then decreasing within penalty 0 while there was no such pattern in penalty 0.01. The difference from 0.01 to 0.1 was notable for the proliferation of more significant genes as evidenced by the greater number of lighter colored regions in the left heatmap. Hence, it seems that the penalty weight's influence on the inference of the gene expressions decreases as the weight increases in magnitude. In addition, the penalty weight seemed to have an insignificant impact on the inferred fraction profile as evidenced by the fact that all the fraction profiles were identical (Fig. 04).

Comparison with prior work

Previous work has focused on making use of both bulk and single-cell genetic sequence data but under slightly different contexts, making this study unique in that regard. Prior work chose a similar strategy of hybrid bulk and single-cell data but with DNA-seq rather than RNA-seq (Lei et al., 2020; Malikic et al., 2017; Salehi et al., 2017). Some studies used bulk DNA-seq to infer on single-cell RNA-seq data (McCarthy et al., 2020; Shafighi et al., 2021). Closer in nature to this study was a prior study developing a tool called bMIND which used both bulk and single-cell RNA-seq data rather than DNA-seq (Wang et al., 2021). The goal of bMIND, however, differs from the goal of RADs as the former focused on using paired data from individuals to identify the cell types present within that data. Another similar study developed a deep neural network called Scanden to infer the cellular profiles of tissues (Menden et al., 2020).

Future work

One future area to explore is to incorporate self-contained data pre-processing into RADs. This would further optimize the process of semi-deconvolution by transferring the responsibility of pre-processing a certain pairing of bulk and reference data from the user to the algorithm itself. The algorithm in its current incarnation requires this as a pre-condition. Another future area is to incorporate newer spatial transcriptomic methods which are approaches that profile how gene expression changes across a tissue. Finally, another promising future route is single-nucleus RNA sequencing as using the higher-resolution nuclei in addition to, or in lieu of, the lower-resolution single cells may perform at least as well as single cells, which would likely make nucleus sequencing the better approach for its lighter cost due to the smaller size of nuclei relative to the entire cell. In fact, prior work seems to suggest promise in this area (Lacar et al., 2016; Ding et al., 2020).

References

S. Agrawal, S.A.B.C. Smith, S.G. Tangye, W.A. Sewell, Transitional B cell subsets in human bone marrow, *Clinical and Experimental Immunology*, Volume 174, Issue 1, October 2013, Pages 53–59, https://doi.org/10.1111/cei.12149

Cirri, P., & Chiarugi, P. (2011). Cancer associated fibroblasts: the dark side of the coin. American journal of cancer research, 1(4), 482–497.

Chu, V.T. and Berek, C. (2013), The establishment of the plasma cell survival niche in the bone marrow. Immunol Rev, 251: 177-188. https://doi.org/10.1111/imr.12011

Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., ... & Levin, J. Z. (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nature biotechnology, 38(6), 737-746.

Galán-Díez, M., Cuesta-Domínguez, Á., & Kousteni, S. (2018). The bone marrow microenvironment in health and myeloid malignancy. *Cold Spring Harbor perspectives in medicine*, *8*(7), a031328.

Kuipers, J., Jahn, K., & Beerenwinkel, N. (2017). Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1867*(2), 127-138.

Lacar, B., Linker, S. B., Jaeger, B. N., Krishnaswami, S. R., Barron, J. J., Kelder, M. J. E., Parylak, S. L., Paquola, A. C. M., Venepally, P., Novotny, M., O'Connor, C., Fitzpatrick, C., Erwin, J. A., Hsu, J. Y., Husband, D., McConnell, M. J., Lasken, R., & Gage, F. H. (2016). Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nature communications*, *7*, 11022. https://doi.org/10.1038/ncomms11022

Lei, H., Guo, X. A., Tao, Y., Ding, K., Fu, X., Oesterreich, S., Lee, A.V., & Schwartz, R. (2022). Semi-deconvolution of bulk and single-cell RNA-seq data with application to metastatic progression in breast cancer. *Bioinformatics*, *38*(Supplement_1), i386-i394. https://doi.org/10.1093/bioinformatics/btac262

Lei, H., Lyu, B., Gertz, E. M., Schäffer, A. A., Shi, X., Wu, K., ... & Schwartz, R. (2020). Tumor copy number deconvolution integrating bulk and single-cell sequencing data. Journal of Computational Biology, 27(4), 565-598.

Lim, B., Lin, Y., & Navin, N. (2020). Advancing cancer research and medicine with single-cell genomics. *Cancer cell*, *37*(4), 456-470.

Malikic, S., Jahn, K., Kuipers, J., Sahinalp, C., & Beerenwinkel, N. (2017). Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. bioRxiv. Nature Communications, 10(2750), 1-12.

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., ... & Flicek, P. (2023). Ensembl 2023. *Nucleic acids research*, *51*(D1), D933-D941.

McCarthy, D. J., Rostom, R., Huang, Y., Kunz, D. J., Danecek, P., Bonder, M. J., ... & Teichmann, S. A. (2020). Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nature methods*, *17*(4), 414-421.

Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D. S., Kloiber, K., ... & Bonn, S. (2020). Deep learning–based cell composition analysis from tissue expression profiles. *Science advances*, *6*(30), eaba2619.

Naxerova, K., & Jain, R. K. (2015). Using tumour phylogenetics to identify the roots of metastasis in humans. *Nature reviews Clinical oncology*, *12*(5), 258-272.

Salehi, S., Steif, A., Roth, A., Aparicio, S., Bouchard-Côté, A., & Shah, S. P. (2017). ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome biology*, *18*, 1-18.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., … Sherry, S. T. (2022).

Database resources of the national center for biotechnology information. *Nucleic acids research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Shafighi, S. D., Kiełbasa, S. M., Sepúlveda-Yáñez, J., Monajemi, R., Cats, D., Mei, H., ... & Szczurek, E. (2021). CACTUS: integrating clonal architecture with genomic clustering and transcriptome profiling of single tumor cells. *Genome medicine*, *13*, 1-16.

The Uniprot Consortium. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523-D531.

Wang, J., Roeder, K., & Devlin, B. (2021). Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data. *Genome research*, *31*(10), 1807-1818.

Wu, S. Z., Roden, D. L., Wang, C., Holliday, H., Harvey, K., Cazet, A. S., Murphy, K. J., Pereira, B., Al-Eryani, G., Bartonicek, N., Hou, R., Torpy, J. R., Junankar, S., Chan, C. L., Lam, C. E., Hui, M. N., Gluch, L., Beith, J., Parker, A., Robbins, E., … Swarbrick, A. (2020). Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. *The EMBO journal*, *39*(19), e104063. https://doi.org/10.15252/embj.2019104063

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., ... & Kasprzyk, A. (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database*, *2011*, bar026.