Bulk Deconvolution Benchmarking

Simon Lee, Mukund Varma

Celsius Therapeutics

Abstract

Bulk RNA-sequencing (RNA-seq) has revolutionized the field of transcriptomics by enabling the analysis of gene expression patterns in complex biological samples. However, bulk RNA-seq data often represents a mixture of multiple cell types, making it challenging to dissect the gene expression profiles of individual cell populations within the sample. Bulk RNA-seq deconvolution is a computational approach that aims to overcome this challenge by estimating the relative proportions and gene expression profiles of different cell types. This paper provides an overview of bulk RNA-seq deconvolution, including its principles, methodologies, applications, and limitations. We discuss the significance of deconvolution in unraveling cellular heterogeneity, to advance personalized medicine. Additionally, we explore various computational algorithms commonly used for bulk RNA-seq deconvolution and highlight emerging trends and future directions in this rapidly evolving field.

Introduction

Inflammatory Bowel Disease (IBD)

Inflammatory bowel disease (IBD) is a chronic condition that involves inflammation of the digestive tract. The two main types of IBD are Ulcerative Colitis and Crohn's disease, which can cause symptoms such as diarrhea, abdominal pain, and fatigue. IBD is a complex and challenging condition to treat because its exact cause is unknown, and it can vary in severity and symptoms from person to person (Silva et. al., 2016).

One of the biggest challenges in treating IBD is the fact that there is no cure for the condition. While medications can help manage symptoms and reduce inflammation, they cannot eliminate the disease. Additionally, IBD can cause complications such as strictures, fistulas, and abscesses, which can require surgery to correct. Another challenge in treating IBD is that the disease can be unpredictable and can flare up at any time. This can make it difficult for doctors to find the right treatment plan, as what works for one person may not work for another.

Despite these challenges, advances in research have led to new therapies and treatment options for IBD. In 2020, the Crohn's and Colitis Foundation launched the IBD summit initiative to help discover biomarkers for both disease onset, and drug discovery purposes (Honig et al., 2020). The purpose of launching this initiative was because only 30-40% of patients respond to primary nonresponse and secondary response biologics which only allows a subset of patients to receive proper treatment. To help advance research, the Crohn's and Colitis Foundation has also introduced IBD Plexus, which is a research program that aims to create a comprehensive database of clinical and biological data from people with inflammatory bowel disease (IBD). The aim of this program is to accelerate the development of personalized therapies for people with IBD. The program collects data from patients, clinicians, and researchers across the United States to help improve the understanding and treatment of IBD. The data collected includes clinical information such as disease course, treatment history, and quality of life, as well as biological information such as genomic and microbiome data. The IBD Plexus program also

provides resources and tools for researchers to access and analyze the data, with the goal of creating a new treatment opportunity with those who struggle from IBD.

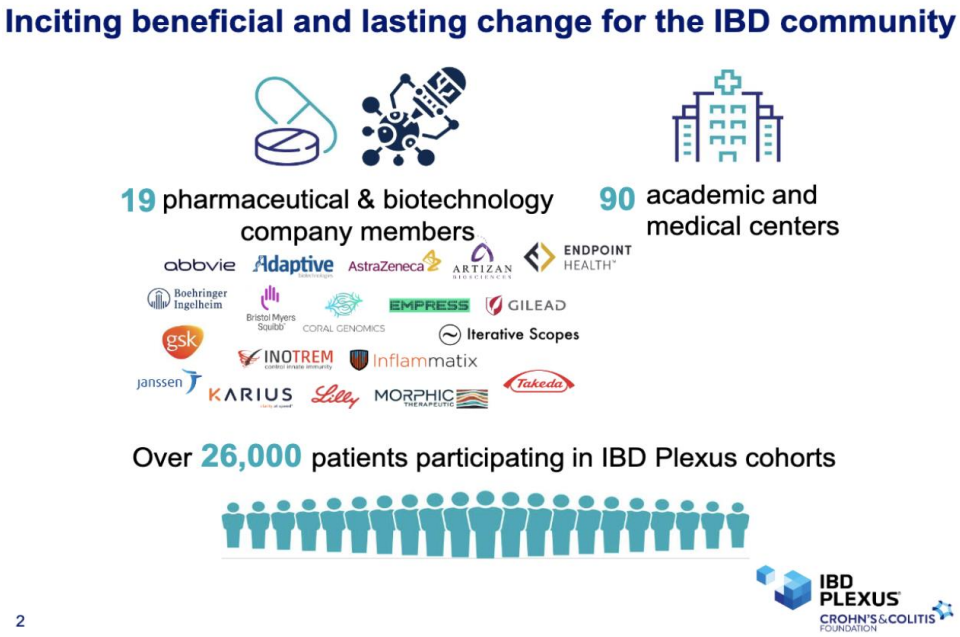


Figure 1. A visual overview of the IBD Plexus Cohort. With many collaborators from both Academia and Industry, there is massive promise contained in this comprehensive database collected by the Crohn's and Colitis Foundation.

At *Celsius Therapeutics*, we have begun our own precision medicine program in hopes of finding promising results that will help us address some of the uncertainties and questions regarding IBD. Among some of these research questions, we have had a prominent level of interest in integrating Bulk RNA-sequencing data in parallel with our single cell platform to help us validate and gain brand new insights from these genetic datasets. In the proceeding sections we will describe Bulk RNA-seq, some of its challenges, and ways to overcome these challenges via computational deconvolution methods.

The Significance and Challenges of Bulk RNA-seq

Bulk RNA-sequencing (RNA-seq) has emerged as a powerful technique for studying gene expression patterns in biological samples. By providing a comprehensive snapshot of the transcriptome, it allows researchers to unravel the complex interplay of genes and their regulatory networks. Bulk RNA-seq has been widely used in various fields, including molecular biology, genetics, and biomedical research, to investigate gene expression changes associated with development, disease, and treatment responses.

Traditionally, bulk RNA-seq involves extracting RNA from a mixed population of cells or tissues and generating sequencing libraries for high-throughput sequencing. The resulting sequencing data provides a snapshot of the average gene expression levels across all cells within

the sample. However, biological samples are often composed of multiple cell types with distinct gene expression profiles, leading to a concept known as *cellular heterogeneity.*

Understanding cellular heterogeneity is crucial as it plays a vital role in various biological processes, including tissue development, immune responses, and disease progression. Cell populations with distinct gene expression profiles can drive specific functions and contribute differently to the overall behavior of a tissue or organ. For instance, in complex diseases such as cancer, different cell types within the tumor microenvironment can have diverse roles, influencing tumor growth, immune evasion, and therapeutic responses (Zaitsev et. al., 2022).

Furthermore, cellular heterogeneity can lead to misinterpretation of bulk RNA-seq data when studying diseases or experimental perturbations. Changes in gene expression levels observed at the bulk level might be attributed to a specific condition, while they could be driven by alterations in the proportions or activities of different cell types within the sample. Disentangling the contribution of individual cell types and accurately characterizing their gene expression patterns is critical for obtaining a comprehensive understanding of complex biological systems. Therefore, when using these datasets, it is worth carefully considering these limitations in analysis related tasks.

While we have mentioned some of the downsides of bulk RNA-seq, it is also worth noting some of its main advantages. One of the primary advantages of bulk RNA sequencing is that it allows for the simultaneous analysis of gene expression across thousands of cells, providing a more comprehensive view of the transcriptome of a particular tissue or organ. Bulk RNA sequencing is also less expensive and less technically challenging than scRNA-seq, which can be an important consideration for researchers with limited resources or experience in molecular biology techniques. Additionally, bulk RNA sequencing can provide higher sequencing depth, which can improve the accuracy of transcript quantification and the detection of lowly expressed genes. Being able to get this sequencing depth allows us to study granular cell types that we cannot traditionally study in single cell technologies.

Motivation for Bulk RNA-seq Deconvolution

In the previous section, we discussed some of the limitations posed by bulk RNA-seq which have motivated the development of bulk RNA-seq deconvolution methods. These approaches offer several key motivations and benefits for unraveling the cellular composition and gene expression profiles within mixed samples.

1. **Enhanced Resolution of Cell Type-Specific Gene Expression**: Bulk RNA-seq deconvolution allows researchers to gain a more detailed understanding of gene expression patterns within individual cell populations. By estimating the gene expression profiles specific to each cell type, deconvolution provides enhanced resolution compared to analyzing the average expression across all cells. This resolution enables the identification of cell type-specific biomarkers, regulatory pathways, and functional characteristics that might otherwise be obscured in bulk data analysis.
2. **Dissecting Complex Tissue and Organ Systems:** Many biological systems, such as organs and tissues, consist of intricate mixtures of cell types that collectively contribute

to their overall function. Bulk RNA-seq deconvolution enables researchers to dissect and understand the specific roles played by different cell populations within these complex systems. By unraveling the gene expression profiles of individual cell types, deconvolution provides insights into cellular heterogeneity, cellular interactions, and functional specialization within tissues and organs.

3. **Uncovering Disease Mechanisms:** In the context of disease research, bulk RNA-seq deconvolution holds great promise. By accurately estimating the relative proportions and gene expression profiles of cell types in diseased tissues, deconvolution can reveal alterations specific to different cell populations. This knowledge is invaluable for deciphering disease mechanisms, identifying novel therapeutic targets, and understanding how different cell types contribute to disease progression, treatment response, and therapeutic resistance.
4. **Advancing Personalized Medicine:** Bulk RNA-seq deconvolution has the potential to enhance personalized medicine approaches. By deconvolving bulk RNA-seq data from patient samples, it becomes possible to assess the cellular composition and identify cell type-specific gene expression signatures associated with disease subtypes or treatment response. This information can aid in patient stratification, tailoring treatment strategies, and predicting individual patient outcomes. Deconvolution also offers insights into the effects of specific therapies on different cell types, facilitating the development of targeted interventions.
5. **Integration with Single-Cell RNA-seq:** The integration of bulk RNA-seq deconvolution with single-cell RNA sequencing (scRNA-seq) data represents a promising avenue of research. By combining the strengths of both techniques, researchers can leverage the high-resolution cell type information from scRNA-seq and the transcriptome-wide coverage of bulk RNA-seq. This integration enables a more comprehensive understanding of cellular heterogeneity, cross-validation of results, and improved accuracy in estimating cell type proportions and gene expression profiles.

Overall, the motivation behind bulk RNA-seq deconvolution is to overcome the challenges posed by cellular heterogeneity and gain deeper insights into complex biological systems, disease processes, and personalized medicine. By accurately estimating cell type-specific gene expression profiles, deconvolution methods empower researchers to unlock the hidden information within bulk RNA-seq data and advance our understanding of diverse biological phenomena.

In the following sections, we will delve into the principles, methodologies, and computational algorithms employed in bulk RNA-seq deconvolution. We will explore their applications in diverse research areas and discuss the limitations and future directions of this field.

The Principles of Bulk Deconvolution

Estimating Cell Type Proportions

A fundamental aspect of bulk RNA-seq deconvolution is the estimation of cell type proportions within a mixed sample. Accurate quantification of the relative abundance of different cell populations is essential for understanding their contributions to gene expression profiles and

unraveling the cellular heterogeneity. A common analogy to understand this process conceptually is trying to quantify the individual fruits that make up an arbitrary smoothie.

Various computational approaches have been developed to estimate cell type proportions from bulk RNA-seq data. These methods leverage different strategies, including reference-based approaches and signature-based approaches. Here, we provide an overview of these approaches:

1. **Reference-Based Approaches:** Reference-based methods estimate cell type proportions by comparing the gene expression profiles of bulk RNA-seq data to pre-existing reference datasets containing known cell type information. These reference datasets typically comprise gene expression profiles of pure cell types obtained from single-cell RNA-seq or sorted cell populations. By measuring the similarity between the bulk RNA-seq data and the reference profiles, reference-based methods infer the proportions of different cell types within the sample.
2. **Signature-Based Approaches:** Signature-based methods leverage gene expression signatures or marker genes associated with specific cell types to estimate their proportions in bulk RNA-seq data. These methods typically require a predefined set of cell type-specific genes or signatures obtained from independent studies or databases. By quantifying the expression levels of these signature genes in the bulk RNA-seq data, signature-based approaches estimate the relative abundance of each cell type.

It is important to note that each approach has its advantages and limitations. Reference-based methods provide accurate estimates when the reference dataset closely resembles the cell types present in the sample, but they may struggle with novel or rare cell types not represented in the reference. Signature-based methods offer flexibility and can be tailored to specific cell types of interest but require well-defined signature genes.

Additionally, factors such as technical variability, and noise in the data can affect the accuracy of cell type proportion estimation. Preprocessing steps, including various normalization techniques, are often performed to mitigate these issues and improve the reliability of deconvolution results.

Deconvolution Algorithms

Bulk RNA-seq deconvolution methods employ various computational algorithms to estimate cell type-specific gene expression profiles. In this section, we provide our benchmarking framework of different deconvolution algorithms discussed in the wider literature survey (White et al. 2022). Before we dive into the algorithms themselves, we first want to present some challenges found in these types of studies.

Comparing the performance of different deconvolution algorithms is challenging due to the lack of ground truth cell type proportions in real-world datasets. Performance evaluation is typically done using simulated datasets (pseudo-bulk) or by comparing deconvolution results with independent experimental validation (flow cytometry, histopathology). Additionally, the choice of deconvolution algorithm depends on the specific research question, the available reference datasets, and the computational resources available which makes the reproducibility of results a significant challenge.

Another challenge typically found in the literature is that each deconvolution method is deemed to be the "state of the art" (SOTA) but has wide variability depending on two factors. One is the tissue being studied and the cell types associated to that tissue, and the second is the gene signature/reference being supplied to these algorithms. Algorithms can have varying performance where it may predict one cell type particularly well but suffer in another (Fig. 1). One reason for this is that the gene expression signatures/references that are used are the greatest determinant of accuracy in current methods for bulk deconvolution (Sturm et al., 2019; Vallania et al., 2018).

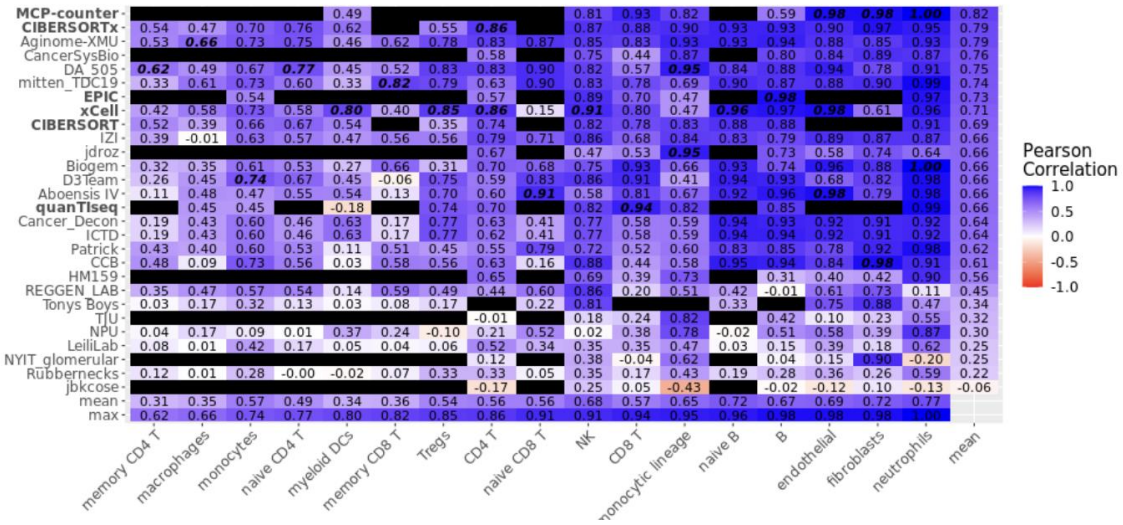


Figure 2. The difference in performance seen across cell types across methods scored on the Pearson Correlation Coefficient. (White et al. 2022)

A final significant challenge in benchmarking has been the unclear guidelines and the lack of a standard metric to score these methods by. Across methods seen in the literature, we see many statistical metrics being taken (e.g., RMSE, Pearson Correlation, Spearman Correlation, R-squared, Mutual information, etc.) to assess the performance of these methods. For reasons not understood, this could bias a method for a particular study which may show optimal performance in the tissue being studied. Therefore, these are all considerations we took when performing our benchmarking study.

In terms of the methods selected for our study, we decided to proceed with taking methods which rely on different computational techniques and those developed in the Python programming language. We briefly describe each algorithm in the following subsections. And in our benchmarking study, we assess the performance of these methods on Pseudo-bulk data from IBD samples coming from proprietary sources (ATAP, Cleveland, LMU) as well as a paired single cell and bulk RNA-seq experiment done in house at *Celsius Therapeutics*.

Kassandra

Kassandra is a decision-tree machine learning based deconvolution method developed by the Boston Gene group (Zaitsev et. al., 2022) and was designed to accurately calculate the proportion of different cell subsets by determining the RNA fraction per cell type from RNA-seq within noncancerous and cancerous tissues. It is a signature-based approach requiring a well-defined gene set to infer the proportions of the cell types of interest. In addition to the gene set, it also considers the hierarchal structure of the cell types, so it does not overpredict or underpredict a parent or child cell type.

Algorithmically this method utilizes the Light Gradient Boosting Decision trees where they fit individual trees for each cell type of interest. It was trained on a diverse range of GEO database bulk samples where an additional artificial mixture is generated to account for more granular cell types and the inclusion of cell specific technical noise. Following the introduction of technical noise into the artificial tumor transcriptomes, the first training stage involved training the cell-type models on TPM-calibrated expression values. The goal was to obtain an RNA fraction per cell type from RNA-seq. In the second stage, the training data comprised gene expression along with the predicted RNA percentages per cell type obtained in the first stage. This stepwise training method facilitated the model's adaptation by leveraging information from other cell types and subtypes for their corresponding models. Furthermore, this approach allowed for the hierarchical utilization of all datasets for artificial transcriptomes.

In terms of limitations, this method does require a high degree of computing power to obtain the pre-trained model from their publication. Additionally, while its main advantages are that it is robust to many tissue types, it does mention the challenge of introducing new and rare cell subtypes that may be of interest (e.g., A lack of a neutrophil signature). We found this to be an apparent issue in our study where there were no considerations to the Stroma or Epithelium compartment which made it difficult to benchmark with other methods.

CIBERSORT (Cell-type Identification by Estimating Relative Subsets of RNA Transcripts) - SVR

CIBERSORT is a regression-based deconvolution method developed by Stanford University (Newman et al. 2015) and was designed to accurately estimate the proportions of multiple cell types simultaneously in each tissue sample. It is a reference-based approach which takes in a gene expression matrix of the cell types of interest obtained from single-cell RNA-seq. On top of estimating the proportions of the relative cell types, it was deemed to be one of the more popular methods due to its ability to build estimated gene expression matrices by minimizing gene expression variance within the same cell type and by maximizing variance between cell types. So, unlike other methods, it provides additional functionality to traditional deconvolution methods.

This algorithm uses a Support Vector Regression algorithm trained on a reference gene expression matrix containing known cell type-specific gene expression signatures. The algorithm generates a linear kernel model to estimate the cell type proportions in the tissue of interest.

Regression algorithms are well-suited for bulk RNA-seq deconvolution because they can model the relationships between gene expression profiles and the proportion of cell types in a mixed sample. Specifically, SVR can learn the gene expression signatures specific to individual cell types and use them to infer the relative proportions of these cell types in a complex mixture. In contrast to other deconvolution methods that rely on reference gene expression profiles or gene sets, regression-based methods do not require prior knowledge of the cell types present in the sample. Instead, they can identify the specific gene expression patterns that are associated with each cell type, allowing for more accurate and comprehensive characterization of complex biological systems.

While SVR does have many advantages, it does have some key limitations. Firstly, they assume a linear relationship between gene expression and cell proportions, which may not be the case in some biological contexts. Additionally, they rely on a quality reference dataset and particularly struggle when trying to accurately identify and quantify rare or novel cell types, leading to inaccurate results. Furthermore, this algorithm does not account for the heterogeneity of individual cells within a given cell type, which can lead to overestimation or underestimation of cell proportions. And lastly this method may be sensitive to technical variability such as batch effects, noise, or differences in sequencing depth, which does not make this algorithm robust.

Cellanneal

Cellanneal is an optimization-based deconvolution method developed at the Weizmann Institute of Science (Buchauer et al., 2021) that takes in a reference gene expression matrix as input and was designed to maximize the correlation by considering ranks instead of absolute data values. This approach allows each gene to influence the optimization result, addressing an issue found in regression algorithms where results are influenced by highly expressed genes. This issue found in regression-based algorithms arises due to the skewed nature of mRNA copy number distributions, which range from less than 1 to more than 10,000 average mRNA copies per cell (Li et al., 2016; Schwanhäusser et al., 2011).

This algorithm uses a probabilistic optimization technique called simulated annealing to find the optimal solution of a given function. Simulated annealing is the process of slowly decreasing the probability of accepting worse solutions as the solution space is explored. So, in the cellanneal pipeline, it generates an artificial mixture vector and optimizes this mixture against the provided bulk gene expression vector using the Spearman rank correlation coefficients. And, with the guidance of the optimization function, it converges on a solution to provide our deconvolution results.

Like SVR, there are some limitations surrounding the sensitivity of the reference gene expression profiles being provided to the algorithm. In this method, the representativeness of reference datasets will affect performance due to differences in experimental conditions and cell type heterogeneity across samples. Furthermore, an essential factor to consider is the relationship between the reference dataset's size and the performance across different methods, which will be investigated in more detail in the following sections. It is worth mentioning that Cellanneal exhibits the slowest computational speed among the three evaluated algorithms due to its optimization strategy, potentially leading to longer execution times.

Data

Pseudo-bulk Experiment Data

Pseudo-bulk data refers to a representation of bulk gene expression data that is derived from single-cell RNA sequencing (scRNA-seq) data. In scRNA-seq experiments, individual cells are sequenced to capture their gene expression profiles. Pseudo-bulk data is generated by aggregating the gene expression profiles of cells within the same group or condition, such as a specific tissue, treatment, or cell type. This aggregation allows for the analysis of gene expression at a population level, like traditional bulk RNA sequencing, where gene expression is measured from a mixture of multiple cells. Pseudo-bulk data provides a way to leverage the rich information obtained from scRNA-seq experiments while enabling analysis techniques developed for bulk RNA-seq data. It can be useful in studying cellular heterogeneity, identifying differentially expressed genes, and performing downstream analyses that require bulk-level gene expression information.

We therefore will look at this pseudo-bulk data to conduct some preliminary benchmarking analysis of these three methods. Since the single cell reference and "bulk samples" come from the same source, we should expect to obtain reliable results across all methods. We utilize our collaborations with ATAP, Cleveland Clinic, and LMU to generate this data in house.

Paired Single Cell and Bulk Experimental Data

Previous benchmarking studies have evaluated the performance of deconvolution methods on primarily pseudo-bulk datasets. While this may indicate great performance for a particular method, it does not yield results that are valid when used in a clinical setting. One reason for this is because pseudo-bulk data comes directly from the single cell reference which in turn results in good performance. Therefore, in our second study, we want to perform an experiment on paired single cell and bulk experimental data coming from and not from the same pinch to assess both the method and the bulk deconvolution results in a more realistic simulation. By conducting a paired single cell and bulk experiment, we also can assess how methods perform in the presence of batch effects that are seen strictly in the scRNA-seq that are not seen in our bulk datasets.

Batch effects in single-cell RNA-seq refer to systematic variations in gene expression that arise due to technical differences between experimental batches or sample processing procedures. These technical factors can include variations in library preparation, sequencing platforms, reagent lots, or experimental conditions. Batch effects can introduce artificial variability and confound the biological signal, making it challenging to accurately identify and interpret true biological differences across cells or conditions. They can lead to clusters or subpopulations of cells that are primarily driven by the batch rather than the biological condition of interest, undermining the ability to draw meaningful conclusions from the data. Therefore, performing a deconvolution experiment on a paired single cell and bulk experiment is crucial to assess the performance of these methods in a simulated clinical setting.

Results & Discussion

Deconvolution Results

In this section, we cover the various analyses conducted as part of our bulk deconvolution benchmarking, which provided insights into the accuracy of our results. We explored a series of different analyses that tested both the inputs to our algorithm, such as gene signatures, and the algorithm itself. The subsequent subsections provide a brief description of each experiment along with our results for pedagogical purposes.

Gene Signatures

Based on recent findings in the published literature (Aubin, R. G et al., 2023), there is an increasing emphasis on the importance of gene signatures and their critical role in obtaining accurate bulk deconvolution results. In our analysis, gene signatures are defined as a gene by cell type expression matrix that guides our algorithms in predicting proportions. However, a significant challenge arises from the lack of guidance and instructions provided in published research on how to prepare these crucial algorithm inputs. Additionally, gene signatures are rarely included in public repositories, further exacerbating the issue. To address this gap, we have conducted a series of compartment-level experiments in the following subsections to assess the robustness and sensitivity of both the gene signatures and the algorithms used.

Gene Signature Size

One of the first experiments we conducted in-depth focused on the effect of signature size on the overall deconvolution results. Each compartment in our study has a unique set of marker genes. However, we were interested in exploring whether incorporating shared overlapping gene information could improve the results. To investigate this, we generated gene signatures and selected the top N genes from each compartment for the analysis. Starting with a signature size of 20 genes, we gradually increased the number of genes and re-ran the pipeline to compare the outcomes. We went as high as 100 genes to assess the impact.

Figure 2 presents the results of this analysis, indicating a slight improvement in performance based on the Spearman correlation coefficient. We also applied this analysis to our pseudobulk datasets (ATAP, LMU, Cleveland) to assess the consistency across different datasets. Preliminary results demonstrate that there was not a significant difference between using a small gene signature size and a larger one, with minor exceptions observed in the case of cellanneal, which showed a marginal increase in performance. We present our findings through a line plot, where each line represents a compartment, along with a scatter plot illustrating the true vs. predicted values for two instances, particularly highlighting the cellanneal algorithm. Moving forward, we consider the fitted linear line, which indicates the quality of data fit. In essence, a slope above 1.0 indicates underprediction of the compartment, while a slope below 1.0 indicates overprediction. This assessment approach will be utilized in future analyses to evaluate performance.

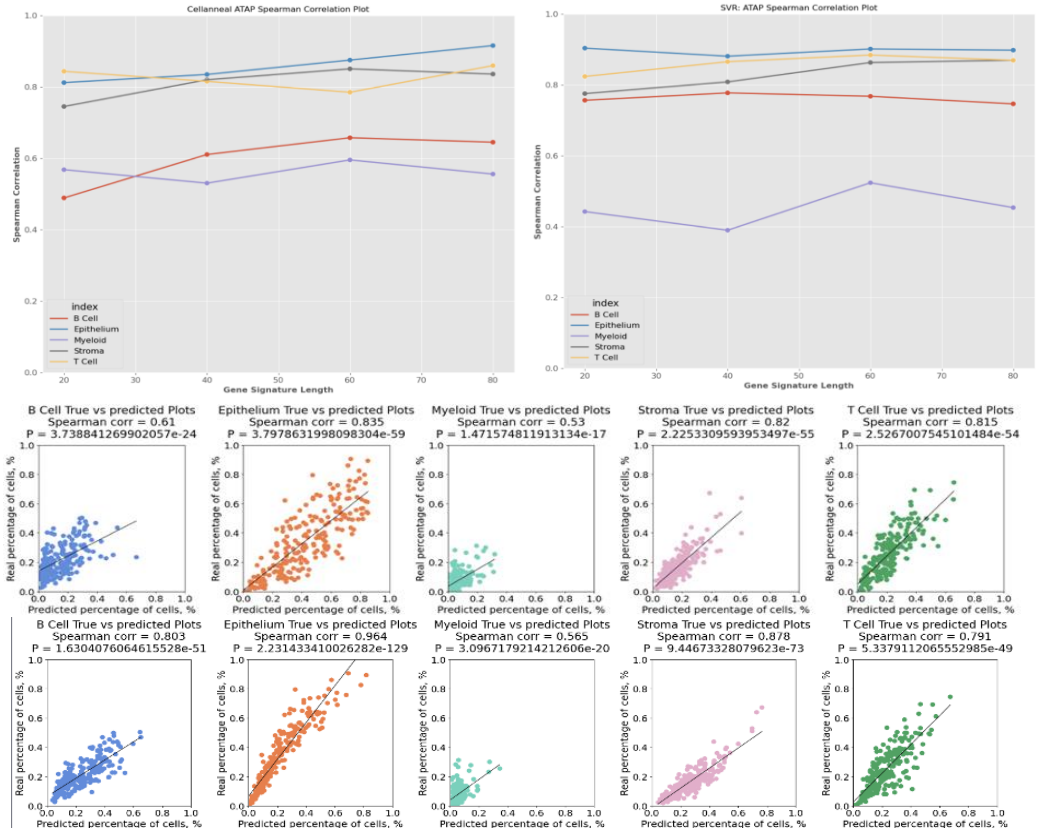


Figure 3. Gene Size Experiment. In the top line plots, we observe the performance of the SVR and cellanneal algorithms at the measured gene sizes based on the Spearman Correlation Coefficient. Each individual line represents a different compartment. Overall, there is a marginal increase in performance at best. Below, we have two sets of scatter plots measured on the ATAP dataset to show a more visual representation of the line plots. The top true vs predicted plots show the performance at a gene signature size of 20, and the bottom at 80 genes. Still, the performance increase is marginal.

Benchmarking Different Preprocessed Data

Previous studies have reported that pre-processing of input data in deconvolution impacts the deconvolution performance (Jew et al., 2020; Avila Cobos et al., 2020), but it is yet unclear which methods should be used for each deconvolution method selected. Standard preprocessing of bulk and single-cell data includes the transformation and normalization of the expression matrices. Transformation methods account for the mean-variance dependencies and extreme count values in a dataset. This is a crucial pre-processing step for most downstream analyses of both single-cell and bulk RNA-seq data. So, in this second exploration, we looked into seeing how to preprocess the bulk samples to achieve the best results. We explored taking a log normalization of our data as well as a logcpm (Median of Ratios Normalization) as well as inputting raw counts of our bulk samples to our algorithms.

In brief, there was not a distinct difference in the different preprocessing steps we experimented with. Though there was some variation in the Spearman Correlation coefficient among the

preprocessing methods and across different datasets, it was not significant enough to explore further. Therefore, as it is a standard practice in many bulk RNA-seq related research, we have adapted the logcpm (Median of Ratios Normalization) as our way of preparing our bulk samples that we would like to deconvolute.

Compartment Knockout Studies

Our final exploratory analysis of gene signatures aimed to assess the robustness of each algorithm when an entire compartment was removed from the signature. In this study, we designed an experiment to create an "other compartment" that would serve as a category for algorithms to converge on when a particular compartment was knocked out. We wanted to determine whether the algorithms would either converge on an existing compartment or fall back on the "other compartment" when no suitable fit was found. To achieve this, we artificially added two additional columns to our gene signature matrix, representing the "others compartment." These columns consisted of a column of zeros and a column with the gene-wise mean. By summing these two columns after running the pipeline, we obtained the final proportion counts. For the experiment, we systematically knocked out each compartment one by one to benchmark the algorithms.

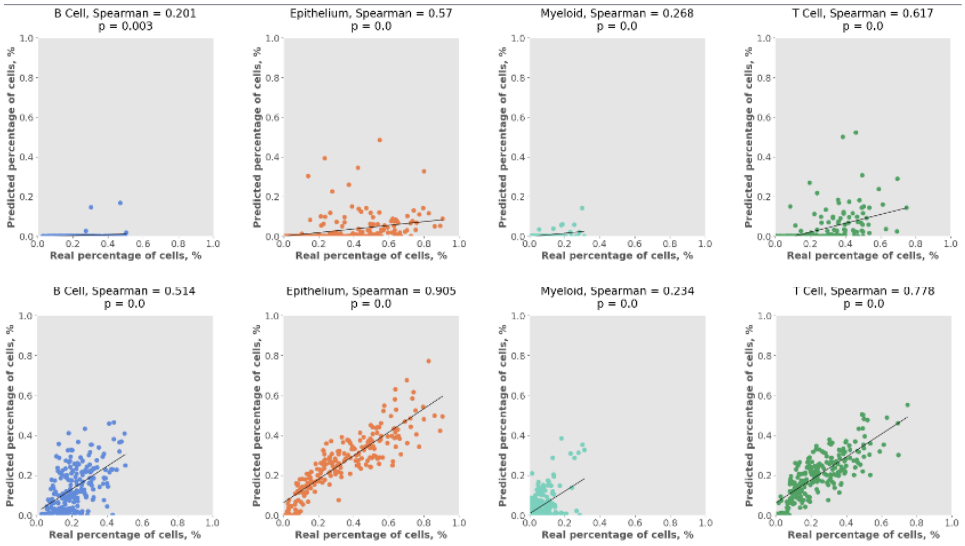


Figure 4. Stromal Compartment Knockout: In these true vs predicted plots, we were particularly interested in assessing the performance of the SVR (top) and Cellanneal (bottom) algorithms when a compartment was completely knocked out. As depicted above SVR tends to "break" when the stromal, epithelial and T cell compartments were knocked out while cellanneal was able to still make proper predictions with marginal decreases in the Spearman correlation.

From our results, we observed that the SVR algorithm struggled to adjust when an entire compartment was knocked out. Specifically, when T cells, Epithelial cells, and Stromal cells were removed, the SVR algorithm exhibited significant difficulties. On the other hand, the cellanneal algorithm, while experiencing a slight reduction in Spearman correlation, appeared to adapt to the knockout and still converged on relatively accurate solutions. Figure 3 provides a clear illustration of this, demonstrating the response when the Stromal cells were knocked out in

the ATAP dataset. It is worth noting that the algorithms performed well when B cells and Myeloid cells were knocked out, possibly due to the specific signature and the complexity of predicting a compartment with low cell proportions from the start. Nevertheless, based on this analysis, we can rule out the SVR algorithm due to its lack of robustness.

Final Algorithm

From our preliminary analysis, we were able to explore the many caveats that are part of these deconvolution algorithms' inputs through various exploratory analyses on the gene signatures. However, we also wanted to work with other datasets, including some that were generated in-house at *Celsius*. Therefore, in our last study, we explored running our bulk deconvolution on a paired single-cell and bulk dataset known as LMU 2.0, which was generated from the same pinch. By having such a dataset, we can assess the true capability of the algorithm instead of relying on some artificial pseudo-bulk proprietary dataset.

In our last study, we were able to identify that Cellanneal was the algorithm that outperformed in our benchmarking experiment. We assessed the algorithms based on the high Spearman Correlation Coefficient (average of 0.781 across compartments), paired with the best linear regression slope closest to 1.0, as our standard for determining the best algorithm. Additionally, we ruled out SVR and Kassandra due to concerns about their reliability and robustness in our compartment knockout study. Based on our comprehensive study, we have identified that Cellanneal is the algorithm we will move forward with in our precision medicine efforts.

Conclusion

Implications and Future Perspectives

In our pursuit of understanding Bulk RNA-seq deconvolution, we recognize that like any research endeavor, there were additional avenues we had planned to explore but were not able to fully investigate in our analysis. Specifically, we had set our sights on delving deeper into the impact of batch effects on deconvolution results, as well as perform a patient stratification using the IBD Plexus Dataset. Unfortunately, unforeseen circumstances hindered our ability to proceed with these aspects of the study.

The investigation of batch effects and their influence on deconvolution outcomes holds significant promise in enhancing the reliability and accuracy of our results. Recognizing and mitigating batch effects is crucial for ensuring the robustness and generalizability of deconvolution techniques, particularly when dealing with diverse datasets from different sources or experiments. While we were unable to embark on this specific exploration, we hope that future iterations of this research, undertaken by other investigators, will unravel these intricacies and provide invaluable insights into the role of batch effects in Bulk RNA-seq deconvolution.

Moreover, an intriguing prospect that we had planned for the study was the application of our algorithms and methodology to the IBD Plexus dataset as part of our precision medicine efforts. This dataset presented a unique opportunity to elucidate the cellular heterogeneity underlying inflammatory bowel diseases and potentially unveil novel biomarkers or therapeutic targets.

Regrettably, due to unforeseen circumstances beyond our control, we were unable to execute this phase of the research.

Summary of Bulk RNA-seq Deconvolution

In this comprehensive study, we embarked on a thorough exploration of the current landscape of Bulk RNA-seq deconvolution, seeking to unravel the intricacies of publicly available algorithms and pioneering benchmarking in novel scenarios that had not been previously addressed in the published literature. Our concerted efforts were aimed at shedding light on the strengths and limitations of the algorithms to provide valuable insights for the precision medicine domain.

While we limited our benchmarking to three Python-based algorithms, the framework we meticulously constructed for this study lays the groundwork for seamless integration of new methods and allows for the future incorporation of R-based algorithms, which are commonly encountered in the literature. This adaptability ensures that our research will remain relevant and extendable as new approaches emerge in this rapidly evolving field.

Through our rigorous analysis, we successfully identified an algorithm that demonstrated superior performance and efficacy for our precision medicine application. We emphasize, however, that this selection is context-specific, and different tissues or more refined signatures might necessitate the utilization of alternate algorithms. The diversity of biological contexts and research objectives in precision medicine mandates a nuanced and individualized approach to algorithm selection. So, in the context of our problem, the algorithm, Cellanneal, was the optimal choice for our present precision medicine endeavors.

Code Availability

All code can be accessed https://github.com/celsiustx/cell-deconvolution.

References

1. Silva, F. A., Rodrigues, B. L., Ayrizono, M. L., & Leal, R. F. (2016). The Immunological Basis of Inflammatory Bowel Disease. Gastroenterology Research and Practice, 2016, 2097274. https://doi.org/10.1155/2016/2097274
2. Honig, G., Heller, C., & Hurtado-Lorenzo, A. (2020). Defining the Path Forward for Biomarkers to Address Unmet Needs in Inflammatory Bowel Diseases. Inflammatory Bowel Diseases, 26(10), 1451-1462. https://doi.org/10.1093/ibd/izaa210
3. Zaitsev, A., Chelushkin, M., Dyikanov, D., Cheremushkin, I., Shpak, B., Nomie, K., Zyrin, V., Nuzhdina, E., Lozinsky, Y., Zotova, A., Degryse, S., Kotlov, N., Baisangurov, A., Shatsky, V., Afenteva, D., Kuznetsov, A., Paul, S. R., Davies, D. L., Reeves, P. M., Lanuti, M., Goldberg, M. F., Tazearslan, C., Chasse, M., Wang, I., Abdou, M., Aslanian, S. M., Andrewes, S., Hsieh, J. J., Ramachandran, A., Lyu, Y., Galkin, I., Svekolkin, V., Cerchietti, L., Poznansky, M. C., Ataullakhanov, R., Fowler, N., & Bagaev, A. (2022). Precise Reconstruction of the TME Using Bulk RNA-Seq and a Machine Learning Algorithm Trained on Artificial Transcriptomes. Cancer Cell, 40(8), 879-894.e16. https://doi.org/10.1016/j.ccell.2022.07.006

4. White, B. S., de Reyniès, A., Newman, A. M., Waterfall, J. J., Lamb, A., Petitprez, F., Valdeolivas, A., Lin, Y., Li, H., Xiao, X., Wang, S., Zheng, F., Yang, W., Yu, R., Guerrero-Gimenez, M. E., Catania, C. A., Lang, B. J., Domanskyi, S.,... Gentles, A. J. (2022). Community assessment of methods to deconvolve cellular composition from bulk gene expression. bioRxiv, 2022.06.03.494221. https://doi.org/10.1101/2022.06.03.494221
5. Sturm, G., Finotello, F., Petitprez, F., Zhang, J.D., Baumbach, J., Fridman, W.H., List, M., and Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* 35, i436–i445. doi:10.1093/bioinformatics/btz363
6. Vallania, F., Tam, A., Lofgren, S., Schaffert, S., Azad, T.D., Bongen, E., Haynes, W., Alsup, M., Alonso, M., Davis, M., et al. (2018). Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nature communications* 9, 4735. doi:10.1038/s41467-018-07242-6.
7. Newman, A., Liu, C., Green, M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**, 453–457 (2015). https://doi.org/10.1038/nmeth.3337
8. Jew, B. et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications* 11, 1971 (2020)
9. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature. Communications* 11, 5650 (2020)
10. Buchauer, L., Itzkovitz, S., cellanneal: A User-Friendly Deconvolution Software for Omics Data. ArXiv:2110.08209
11. Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J. C., Rodig, S., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biology, 17(1), 1–16.
12. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature, 473(7347), 337–342.
13. Aubin, R. G., Montelongo, J., Hu, R., Camara, P. G., (2023). Clustering-independent estimation of cell abundances in bulk tissues using single-cell RNA-seq data. bioRxiv 2023.02.06.527318; doi: https://doi.org/10.1101/2023.02.06.527318

Appendix

I. A Detailed Overview of Inflammatory Bowel Disease

At a high-level view, Inflammatory Bowel Disease (IBD) is characterized by inflammation along the gastrointestinal tract, but the location and extent of the inflammation can vary. While the exact root cause is unknown, it is believed that some imbalance in the gut microbiome disrupts the immune homeostasis of patients. The mucosal barrier in the gastrointestinal tract contains a range of mechanisms that may play a role in the development of IBD.

One such mechanism involves Intestinal Epithelial cells (IECs), which express toll-like receptors (TLRs) that produce pro-inflammatory cytokines to maintain the epithelial layer. However, people with IBD typically have a damaged version of TLR (toll like receptors) receptors, leading to permeable mucosal membranes. Genes such as TLR2 and TLR4 are more active in IBD patients, causing hyper activation and mucosal inflammation.

Another factor at the microbiota level is the imbalance between Treg and Th17 cells, caused by factors such as TGF-β (Transforming Growth Factor) and IL-6. IBD patients have fewer Treg cells and fluctuating Th17 cells, which are regulated by the presence of commensal bacteria, suggesting that environmental factors may be involved. With disruption of the balance and excessive increase in Th17 cells, the development and maintenance of inflammation is suspected.

Macrophages and dendritic cells (DCs) are also responsible for the lack of response to commensal bacteria, affecting gut homeostasis. Important commensal bacterial-driven genes such as IL-22, produced by ILC3s, play a role in protecting against infectious pathogens. Innate immunity response for IBD typically sees an increase in the number of macrophages in the intestinal mucosa, which express many T cells and costimulatory molecules such as CD40, CD80, and CD86, involved in the inflammatory process.

DCs are activated in small numbers but have strongly expressed microbial receptors in Crohn's disease (CD) and ulcerative colitis (UC) patients, leading to over-expression of pro-inflammatory cytokines like IL-6 and IL-12. As such, DCs are fundamental in IBD, responsible for the balance between the tolerance to commensal microorganisms and immune activity by detecting certain molecular structures of the bacteria.

II. An Overview of the Bulk Deconvolution Package

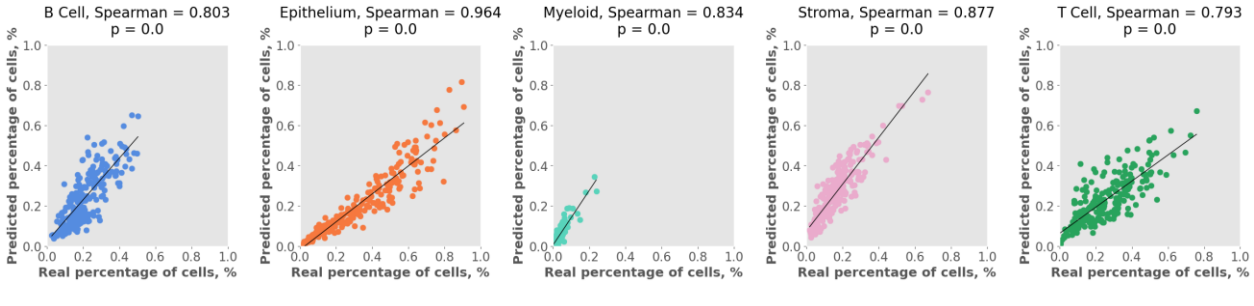
Our analysis was conducted using the bulk deconvolution software package developed by Celsius Therapeutics. The motivation behind the development of this package stems from the lack of clear guidelines for method selection and comprehensive benchmarking across various tissues. In addition to the complexities associated with deconvolution, such as preprocessing input matrices, and selecting appropriate single cell references, the accurate resolution of closely related cell types or cell states remains uncertain.

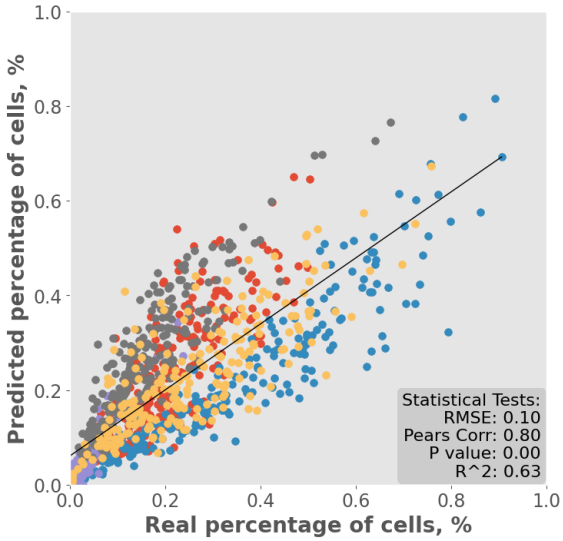
The package therefore helps build a framework where methods can be easily deployed and be assessed on various statistical metrics. Some metrics include RMSE, Residuals, R-squared, and both Pearson and Spearman Correlation coefficients. Additionally, since all results are a proportion, we provide all these statistical metrics in their normalized forms so true high proportions do not feel overemphasized when calculating various error metrics. While our analysis focused on primarily taking the Normalized RMSE (NRMSE) and Normalized Residuals, users are free to use the other available metrics to benchmark performance across methods of interest.

This package also includes many visualization tools to help us understand various aspects of our deconvolution results. For pedagogical purposes, we describe some of the features and their significance to our analysis.

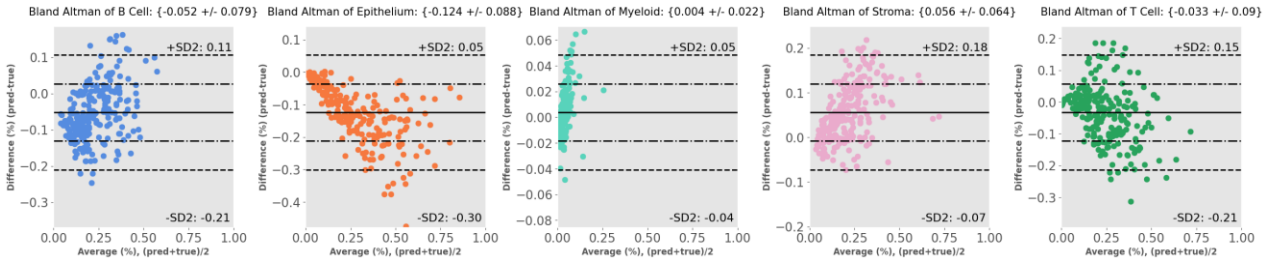
Visualization Methods:

True vs. Predicted Scatterplot: A scatterplot of our "true" vs predicted proportions with the option to fit a linear regression at the cell type or sample level. There is also an option to have one whole scatterplot with all the cell types/samples in one plot. In the cell type/sample level, we also take the Spearman Correlation of the True vs. Predicted vectors and display the p-values to indicate the statistical significance of the deconvolution results. At the whole sample level, we get two additional metrics displayed which are the RMSE and R-squared values.

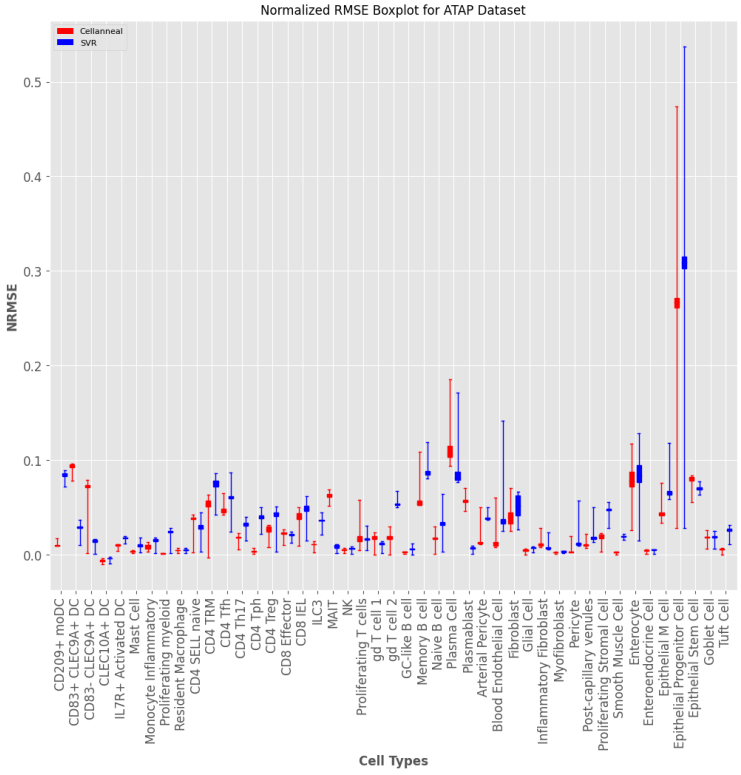




Bland Altman Plots: A Bland-Altman plot is a graphical method used to assess the agreement between two measurement techniques or observers. It involves plotting the residuals between the measurements on the y-axis and the average of the measurements on the x-axis. The plot also includes horizontal lines representing the mean difference and limits of agreement, which indicate the range within which most differences between measurements are expected to fall. These plots help show us our residuals and any technical bias seen in our deconvolution results.

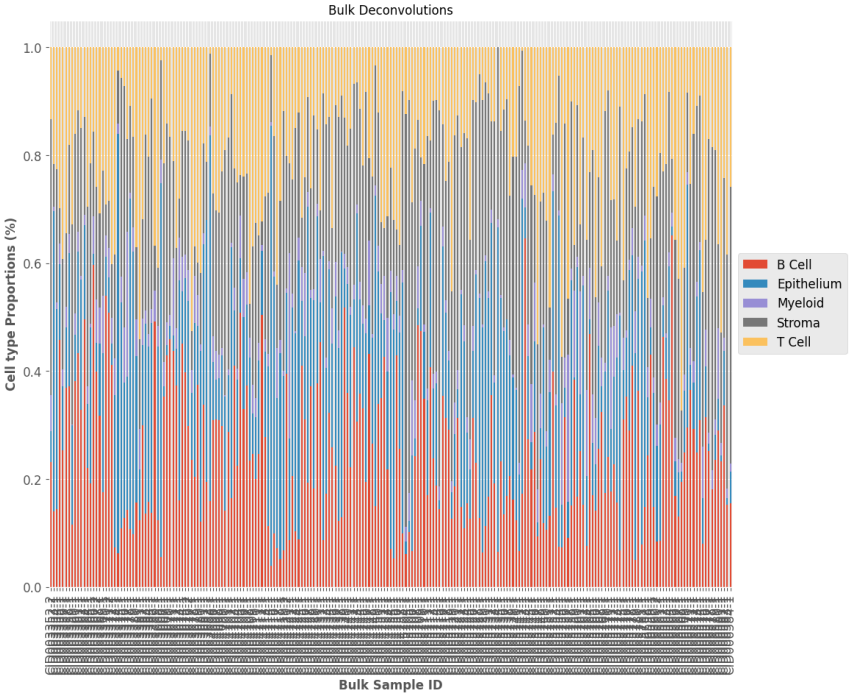


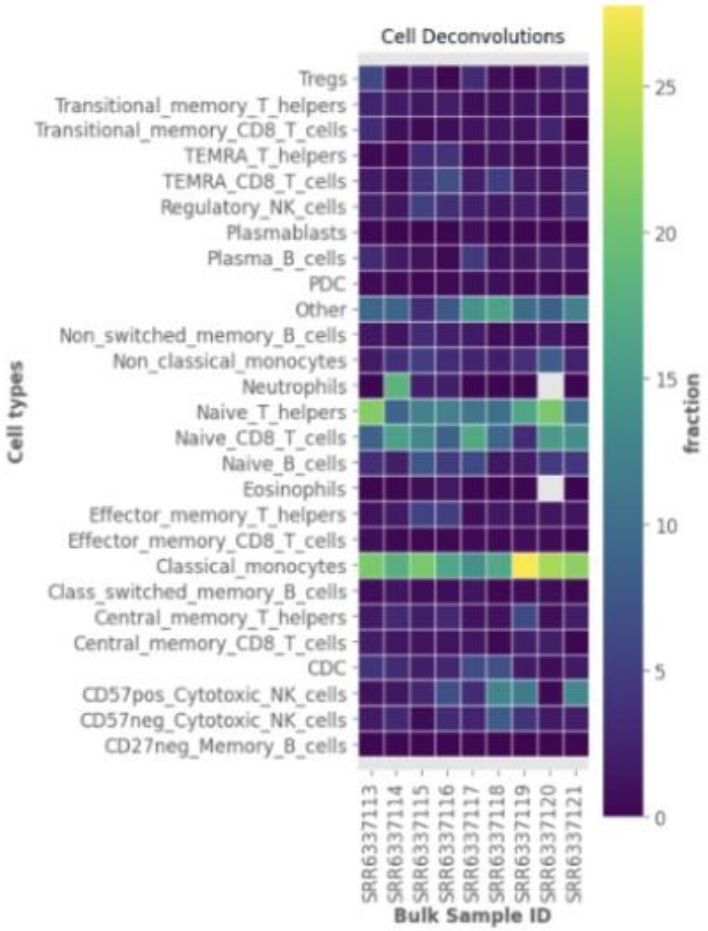
Boxplots: Boxplots provide a concise summary of the distribution of a dataset, allowing for quick comparisons between different cell types or compartments. They effectively display the central tendency, spread, and skewness of the data, making them useful for identifying patterns and detecting potential outliers or differences between groups. In our boxplot module we have cells against normalized RMSE and residuals at a cell type level colored by the methods being benchmarked. Minor changes are required in the module to account for the number of methods being benchmarked.



Heatmaps/Stacked Bar plots

Heatmaps and stacked bar plots are good for visualizing proportions because they provide a clear and intuitive representation of the relative proportions of different our individual predicted cell types/compartments. Heatmaps use color gradients to visually represent the magnitude of proportions, allowing for easy identification of patterns and trends. Stacked bar plots show the cumulative proportions of categories stacked on top of each other, providing a visual comparison of proportions across multiple variables or groups.


Bulk Deconvolutions
Cell type Proportions (%)
1.0
0.8
0.6
0.4
0.2
0.0
B Cell
Epithelium
Myeloid
Stroma
T Cell
Bulk Sample ID


Cell Deconvolutions
Tregs
Transitional_memory_T_helpers
Transitional_memory_CD8_T_cells
TEMRA_T_helpers
TEMRA_CD8_T_cells
Regulatory_NK_cells
Plasmablasts
Plasma_B_cells
PDC
Other
Non_switched_memory_B_cells
Non_classical_monocytes
Neutrophils
Naive_T_helpers
Naive_CD8_T_cells
Naive_B_cells
Eosinophils
Effector_memory_T_helpers
Effector_memory_CD8_T_cells
Classical_monocytes
Class_switched_memory_B_cells
Central_memory_T_helpers
Central_memory_CD8_T_cells
CDC
CD57pos_Cytotoxic_NK_cells
CD57neg_Cytotoxic_NK_cells
CD27neg_Memory_B_cells
Cell types
SRR6337113
SRR6337114
SRR6337115
SRR6337116
SRR6337117
SRR6337118
SRR6337119
SRR6337120
SRR6337121
Bulk Sample ID
25
20
15
10
5
0
fraction

Benchmarking Mathematical Definitions

Spearman Correlation Coefficient:

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

R-squared:

$$R^2 = 1 - \frac{SSE}{SST}$$