ESCUELA DE INGENIERÍA

DE FUENLABRADA

BACHELOR OF SCIENCE IN

BIOMEDICAL ENGINEERING

END OF DEGREE PROJECT

COMPARISON OF EMERGING METHODOLOGIES FOR DECONVOLUTION OF BULK GENETIC EXPRESSION

AUTOR: VICTORIA ANGUIX LÓPEZ

TUTOR: LUIS BOTE CURIEL

COTUTOR: JOSÉ LUIS ROJO ÁLVAREZ

CURSO ACADÉMICO: 2022/2023

Always, always, always believe in yourself.
'Cause if you don't, who will, baby?

Marilyn Monroe (1926-1962)

Acknowledgments

I want to convey my sincere thanks to all the people who have supported me and been by my side during this period of my university education. After a few hard years of effort, a new stage starts full of new opportunities, for which I am very excited. First, thanks to my tutors, Luis and José Luis, who have been committed to the job from the beginning. Without them, this work would not have been possible.

To my university classmates and friends, Juliana and Ana, even though the pandemic robbed us of some of the best years of college and then we drifted apart during Erasmus, we are still as close as we were on that first day of class. Thank you very much for all the support as always. Also to Suke, Alba and Vita, my unconditional companions during the Erasmus. Thank you for making that experience a dream, and above all, thank you for becoming true friends afterwards.

To my parents, Domitila and Arturo, which have unconditionally supported me every time and always have comforting words. Both of you are a referent for me, I owe it all to you, thanks for the effort you make every day so that I can continue with my career. Also, my brother Arturo, my role model, who has helped me along this road both as a 'teacher' and as a brother, with the best advice always. Thank you for all.

Finally, thanks to my couple, Alberto during this last year, you have endured all my negativity and bad times, but you you have always reminded me of my strength and abilities, and once again you were right.

Abstract

Since the advances in the technology of high-throughput sequencing (HTS), scientists are able to sequence several molecules at a time, having huge data sets of genes. Concretely, bulk RNA sequencing (bRNA-seq) measures average gene expression across a population of heterogeneous cells and single-cell RNA sequencing (scRNA-seq) examines gene expression patterns of individual cells. Tissue samples are still routinely processed in bulk due to the complexity of scRNa-seq methodology, which makes it critical to determine the cell-type composition of each sample and the gene expression profile (GEP) of each constituent cell type. By the process of deconvolution, this can be achieved, and concretely, computational deconvolution has a bright future in understanding the underlying mechanisms of many biological processes by its cell type composition. Since 2016, many statistical-based approaches have tried to solve this problem. In addition to all of them, currently, new deep-learning methods have appeared.

Given these circumstances, the main objective of this end-of-degree project (EDP) is to compare the performance of the two more up-to-date methods of bRNA-seq deconvolution on different scenarios and with different types of data, in order to ensure that the evaluation is as thorough as possible.

Accordingly, two algorithms are tested, the SCDC method based on weighted non-negative least squares (W-NNLS) and the Tissue-AdaPtive auto-Encoder (TAPE) method based on Deep Neural Networks (DNNs). They are both tested with two types of data, pseudo bulk with known solution, and real data of human pancreatic islets. For each experiment, different figures and metrics are chosen depending on the type of data, such as can be Concordance Correlation Coefficient (CCC), mean absolute error (MAE), absolute error plots, Bland-Altman plots, Wilcoxon signed-rank test and linear regression algorithms.

After the completion of the EDP, it is concluded that, although SCDC shows better performance on pseudo bulk experiments and TAPE performs badly on them, TAPE is a more robust method that performs better for real data scenarios. However, it has been demonstrated that both methods still have some way to go in terms of deconvolution across protocols. Given the experiments conducted during this EDP, it is difficult to determine which method is better for deconvolution performance, since both of them have their advantages and disadvantages depending on the protocol, size and cell types of the data.

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

10X *10X Genomics Chromium*

Ba *Basophil*

BC *B cell*

bRNA-seq *Bulk RNA sequencing*

CCC *Concordance Correlation Coefficient*

CiCC *Ciliated Columnar Cell of tracheobronchial tree*

CM *Classical Monocyte*

CSV *Comma-Separated Values*

CSx *CIBERSORTx*

DNA *DeoxyriboNucleic Acid*

DNNs *Deep Neural Networks*

DWLS *Dampened Weighted Least Squares*

EC *Endothelial Cell*

EDP *End of Degree Project*

GEP *Gene Expression Profile*

Gr *Granulocyte*

HbA1c *Haemoglobin A1c*

HPC *Hematopoietic Precursor cell*

HTS *High-Throughput Sequencing*

HVG *Highly Variable Gene*

iBC *immature B cell*

LEC *Lung Endothelial Cell*

lpBC *late pro-B cell*

L *Leukocyte*

MAE *Mean Absolute Error*

Ma *Macrophage*

MC *Myeloid Cell*

Mo *Monocyte*

MSC *Mesenchymal Stem Cell*

MSE *Mean Squared Error*

MuSiC *MUlti-Subject SIngle Cell*

MVW *Maximal Variance Weights*

NGS *Next-Generation Sequencing*

NNLS *Non-Negative Least Squares*

PBMC *Peripheral Blood Mononuclear Cell*

RMSE *Root Mean Squared Error*

RNA-seq *RNA sequencing*

RNA *RiboNucleic Acid*

RPKM *Reads Per Kilobase transcript and per million Mapped reads*

scRNA-seq *Single-cell RNA sequencing*

SMSC *Skeletal Muscle Satellite Cell*

SVR *Support Vector Regression*

TAPE *Tissue-AdaPtive auto-Encoder*

TC *T cell*

TPM *Transcripts Per kilobase Million*

UMIFM *Unique Molecular Identifier Fragment per Million counts*

UMI *Unique Molecular Identifier*

W-NNLS *Weighted Non-Negative Least Squares*

Chapter 1

Introduction

This chapter presents the different topics on which the context and motivation of this end-of-degree project (EDP) have been based. Moreover, general and specific objectives are listed. Finally, a short description of the structure of the chapters is summarized.

1.1 Context and motivation

In recent years, we have had many advances in the field of medicine and genomics, even so, there is still a great gap between current treatments and the cure for many genetic diseases, such as cancer. The main feature of these diseases are the changes in gene regulation due to mutations that modify the usual functions of the genes and proteins. This disruption in gene regulation could lead to uncontrolled cell growth and proliferation, which is difficult to control and stop once it begins. Therefore, the root of this issue begins with the understanding of gene mutations, and thus, of human transcriptome [1].

Since the advances in the technology of high-throughput sequencing (HTS), scientists are able to sequence several molecules at a time, having huge data sets of genes. Specifically, focusing on human transcriptome we have two types of Ribonucleic acid (RNA) sequencing (RNA-seq). The first type is bulk RNA sequencing (bRNA-seq), which measures average gene expression across a population of heterogeneous cells, and the second type is single-cell RNA sequencing (scRNA-seq), which examines gene expression patterns of individual cells. Researchers can analyze the transcriptome of different cells within the same tissue type using single-cell technologies. This method is very effective in the study of the heterogeneity of cells in complex tissues, as in tumour heterogeneity research. However, having a more precise tool

such as single-cell sequencing would give a more accurate methodology and quality control [1].

As a result, tissue samples are still routinely processed in bulk, making it critical to determine the cell-type composition of each sample and the gene expression profile (GEP) of each constituent cell type. The term "deconvolution" in this field refers to a process that calculates the proportion of each cell type in a bulk sample, as well as their corresponding cell-type-specific GEPs. Apart from experimental deconvolution, computational deconvolution has a bright future due to innovative approaches emerging in recent years. Since 2016, the first statistical-based approaches were proposed to solve the GEP calculation problem, most of them being based on traditional regression models such as non-negative least squares (NNLS) and support vector regression (SVR) [2]. However, over the past few years, this field has taken a machine learning path. It is well known that the use of big data and machine learning in medicine has revolutionized the landscape of treatments, ushering in a new era of personalized and data-driven healthcare. Furthermore, the use of machine learning in the analysis of transcriptomic data has revolutionized our understanding of cell heterogeneity in different tissues. Machine learning algorithms can analyze transcriptomic data from thousands of genes simultaneously and identify distinct cell types and their states based on their gene expression profiles. Therefore, one of the last approaches to solve what concerns us is the deconvolution of bulk data by using machine learning models [3].

Hence, HTS has allowed scientists to gather vast amounts of reliable datasets with excellent quality, which hold within them the possible solution to genetic diseases such as cancer, they just need to be processed in an appropriate manner. Having the appropriate computational procedure, such as deconvolution, and the proper tools, the idea of joint them together with single-cell and bulk data is a promising hypothesis, but in order for all to work correctly, the use of the proper methods for each type of data is essential, as well as their evaluation. A correct comparison between the last and most developed methods will bring us the best combination for a correct understanding of the underlying mechanisms of many biological processes.

1.2 Objetives

Having in mind the relevance of the combination of both bulk and single-cell data for a nearly absolute understanding of the human transcriptome, the main objective of this EDP is to compare the two more up-to-date algorithms for bRNA-seq deconvolution, having in mind the two groups of existing methods, namely, statistical-based learning method and deep learning based

methods. The main objective is obtained from the following specific objectives:

- To analyze the methods proposed in the current literature in order to find the newest approaches.
- To comprehend of the base of the methods to be tested. Firstly, the SCDC method based on weighted non-negative least squares (W-NNLS) and later, the Tissue-AdaPtive auto-Encoder (TAPE) method based on Deep Neural Networks (DNNs).
- To implement the methods on different datasets in order to compare their functioning in different cases.
- To implement the methods across protocols and different data sizes for evaluating their robustness.

1.3 Methodology

For the development of this EDP, the first step was the search for state-of-the-art methods with more documentation and easy-of-use algorithms. After some research and filtering, the chosen methods to experiment were SCDC and TAPE. The different experiments carried out have needed distinct languages depending on their mode of action. Like most of the statistical-based deconvolution methods, SCDC has a library on R, while TAPE, like most of the deep learning algorithms, has been constructed on Python.

Moreover, each methodology worked with different types of data, so that conversion between types has been necessary. SCDC needs *ExpressionSet* objects with raw read counts as input, while TAPE works with data frames of raw read counts.

Finally, for the evaluation and visualization of the experiments, the metrics of Concordance Correlation Coefficient (CCC) and mean absolute error (MAE) were combined to compare the algorithms, as well as a linear regression model for the evaluation of real data. Also, several 3D representations of the data, box plots, and bar plots helped in the analysis of the experiments.

1.4 Structure of the memory

We next present an overview of the subjects that will be discussed in each chapter of this EDP:

- **Chapter 1: Introduction and objectives.** As an introduction to the EDP, the problem of the study of transcriptomic data and the difficulty of single-cell sequencing are presented. As a solution, some deconvolution algorithms have been developed in recent years and the EDP is oriented to the evaluation and understanding of some of these methods, as a main goal.

- **Chapter 2: Previous concepts.** The main concepts to contextualize the EDP are explained here, from the HGS to the bulk and single-cell sequencing, finishing with cell-type deconvolution methods. The most important methods to date are briefly described, leaving aside for the Methods section the two of them that are used for evaluation purposes.

- **Chapter 3: Data and methods.** First, a detailed description of the data used in the experiments is given, being pseudo-bulk data and real data. Then, an extensive overview of the two methods evaluated is given. Finally, different metrics, plots and tests that will be used for assessing the performance of the methods are explained.

- **Chapter 4: Results**. This section contains the results of the several tests that were out. There are two subsections, separating the type of data used. For each one, a first individual evaluation of the two methods is given, finishing with a comparison between methods by contrasting the different results on metrics given in the figures.

- **Chapter 5: Conclusions and future lines of action.** The final conclusions that have been reached throughout the development of this EDP and several of the possible future lines that this project can cover are presented. Finally, this chapter also describes the competencies from the degree needed to perform these experiments, as the ones acquired during the process.

Chapter 2

Previous concepts

In this chapter, a more exhaustive review of the literature is made, presenting all the necessary concepts to understand the deconvolution algorithms and their different approaches.

2.1 Introduction

HTS techniques, also known as next-generation sequencing (NGS), have become crucial in genomics, epigenomics, and transcriptomics research. While Sanger sequencing has typically been used to clarify sequencing information, NGS methods are capable of sequencing numerous deoxyribonucleic acid (DNA) molecules in parallel, allowing hundreds of millions of DNA molecules to be read at a time. As a consequence of this advantage, HTS may be utilized to generate massive data sets, resulting in more thorough insights into the cellular genomic and transcriptome markers of diverse illnesses and developmental phases. Within HTS technologies, the topic of this EDP relies on RNA-seq, which is used to analyze how the transcriptome changes [4]. We can distinguish two types of RNA-seq, namely, bRNA-seq and scRNA-seq, which are mainly differentiated by the volume and resolution with which they examine gene expression. Apart from that, we can distinguish them by many other characteristics.

On the one hand, the average gene expression across a population of heterogeneous cells is measured via bRNA-seq. RNA from several cell types is collected, pooled, and sequenced in this approach. This yields an average expression profile for the total cell population, which can be valuable for detecting differential expressed genes across tissues, circumstances, or time periods. Nevertheless, bRNA-seq cannot differentiate gene expression variations among individual cells within a population and can obscure uncommon cell populations, minor transcriptional

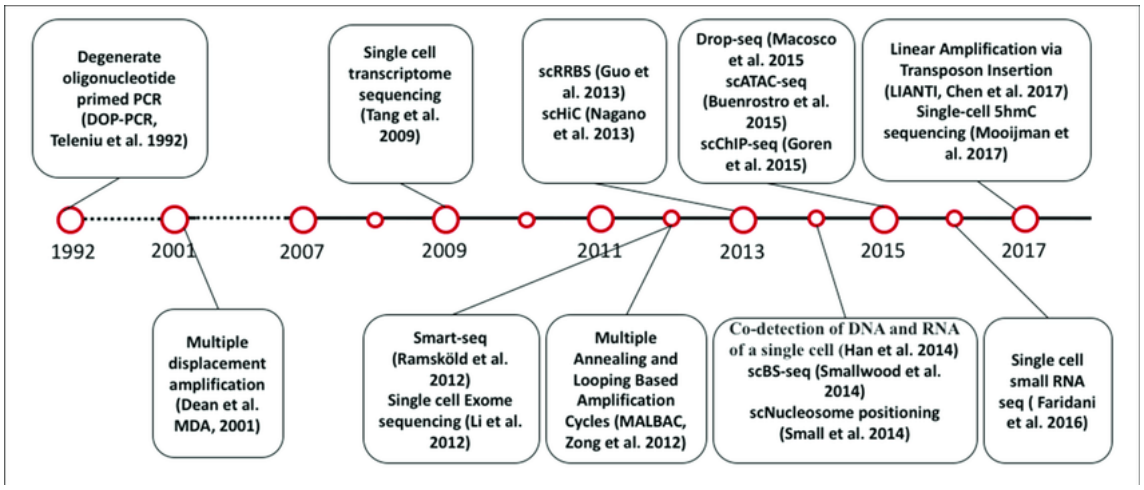


Figure 2.1: Timeline of single cell sequencing methods milestones (from [7], licensed by CC BY 4.0).

discrepancies, or changes in gene expression over time.

This technique is often used to analyze gene expression patterns in diverse tissues, cell types, or experimental circumstances, such as disease states, medication treatments, or time-course research. Also, researchers can detect differentially expressed genes by comparing gene expression patterns through this technique between different environments. This can give insights into the biological processes and molecular pathways implicated in these settings. Furthermore, bRNA-seq may be used to find novel transcripts, isoforms, and non-coding RNAs, as well as annotate and revise current genome annotations. Finally, it can be used to analyze alternative splicing events, which can assist researchers in finding the functional effects of alternative splicing and its control [5].

On the other hand, the gene expression patterns of individual cells originating from homogeneous and heterogeneous populations are examined using scRNA-seq. This method separates single cells, generally through encapsulation or flow cytometry, and then amplifies and sequences the RNA from each cell individually. scRNA-seq is a novel methodology as we can observe in Figure 2.1, [6] from 2009 being one of the first papers published. Researchers may identify cell kinds, states, and subpopulations using this high-resolution method. scRNA-seq can also show cellular heterogeneity and unusual cell types that were previously hidden in bRNA-seq data [5].

This type of sequencing is especially beneficial for researching cellular heterogeneity within tissues, showing diverse cell populations and states that may be obscured in bRNA-seq data. Also, this method is useful for discovering and describing novel or unusual cell types, as well as

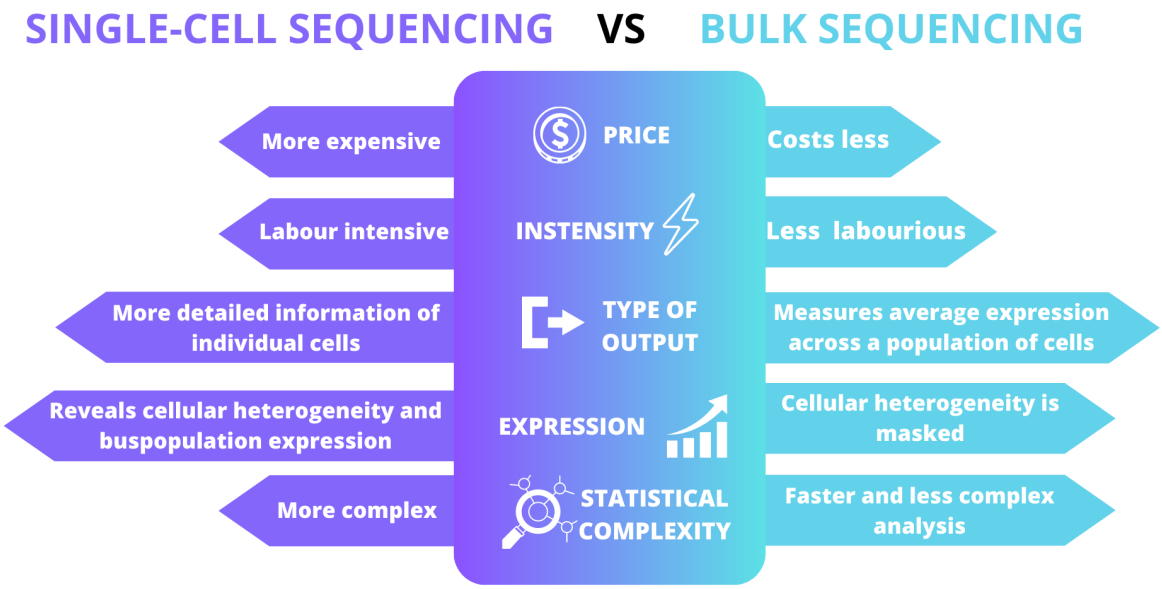


Figure 2.2: Comparison between single cell and bulk sequencing (edited from [10])

refining recognized cell types, by comparing GEPs of individual cells. Additionally, scRNA-seq may be used to explore cellular differentiation, lineage tracing, and developmental trajectories in diverse species, offering insights into the processes driving cellular destiny decisions [5]. As a summary of the properties of both bRNA-seq and scRNA-seq, Figure 2.2 shows some of them.

Regarding scRNA-seq protocols, two popular scRNA-seq technologies are the droplet-based 10X Genomics Chromium technique (10X) and the plate-based Smart-seq2 full-length method. Smart-seq2 uses microtiter plates to isolate mRNA and reverse transcribe it to cDNA for each cell. The number of reads mapped to a gene is used to assess its abundance in each cell, and transcripts per kilobase million (TPM) is a typical metric for normalizing expression. 10X, on the other hand, is a droplet-based scRNA-seq method that allows for genome-wide expression profiling of thousands of cells at once. The number of unique molecular identifiers (UMIs) is thought to be a direct representation of the amount of gene expression. TPM (Smart-seq2) and normalized UMI (10X) are both utilized to find highly variable genes (HVGs), which are frequently employed for cellular phenotype categorization or the identification of novel subpopulations [8]. Apart from them, there are many other protocols less used, for instance, the one described by Klein et al. [9], or the proprietary sequencing protocols from Illumina.

Although the development and innovation of scRNA-seq strategies may gradually lead to a shift from bulk integrative analysis to a detailed investigation of individual cells, bulk strategies complement scRNA-seq approaches in order to obtain whole-system and cell-based perspectives and mechanisms for health and disease. Since bulk profiling approaches cannot correctly detangle cellular heterogeneity, the single-cell investigation is critical for a better understanding of cellular activities and cell-to-cell differences in both fundamental and clinical research. Nevertheless, as compared to traditional bulk techniques, single-cell technologies for specific omics are still in the early stages of development, with poorer capture efficiency and more technical noise. Therefore, combining bulk and single-cell data with deconvolution methods could be a viable way to investigate a huge number of people in a cell-type-specific manner [11].

The concept of "deconvolution" in this field refers to a process that calculates the corresponding cell-type-specific GEPs by the assumption of an algebraic problem described as follows. We have a first matrix $\mathbf{P}$ of size n samples $\times$ k cell types, which represents the proportion of each cell type in a bulk sample, a second matrix $\mathbf{S}$ of size k cell types $\times$ m genes representing a basis or signature matrix (indicates the expression levels of the genes in the corresponding cell types), and a third output matrix $\mathbf{G}$ of size n samples $\times$ m genes that represent the GEP of a specific bulk data. Finally, the problem is defined as:

$$\mathbf{G}_{nxm} = \mathbf{P}_{nxk} \cdot \mathbf{S}_{kxm} \tag{2.1}$$

A simplified example with sizes n six samples, m five genes, and k four cell types and random values is shown in Figure 2.3, but for better representation, a real example of a $\mathbf{P}$ matrix is given on Figure A.1.

In the past few years, many single-cell profile-assisted techniques have emerged to analyze bRNA-seq data. Existing approaches may be broadly classified into two types: statistical learning-based methods and deep learning-based methods. Based on classic regression models such as NNLS and SVR, a number of approaches have been developed, including Bseq [12], MUlti-Subject SIngle Cell (MuSiC) deconvolution [13], CIBERSORTx (CSx) [14], Dampened Weighted Least Squares (DWLS) [15], CDseq [2], Bisque [16], and SCDC [17]. All of these methods, except CDSeq, require pre-selected cell-type-specific GEPs or allocating different weights to different genes based on statistic values, either mean or variance.

Contrary to them, deep learning methods such as Scaden [18] use simulated bulk data for training rather than a pre-defined GEP, and they can automatically extract features from GEP. Despite this advancement, these approaches neglect the operating time cost, which is especially

	Gene1	Gene2	Gene3	Gene4	Gene5
Sample1	0.41	0.67	0.85	0.33	0.79
Sample2	0.45	0.65	0.91	0.23	0.82
Sample3	0.26	0.8	0.83	0.5	0.66
Sample4	0.28	0.45	0.81	0.54	0.91
Sample5	0.27	0.73	0.73	0.59	0.71
Sample6	0.34	0.55	0.74	0.52	0.86

(a)

	Cell1	Cell2	Cell3	Cell4
Sample1	0.68	0.22	0.04	0.06
Sample2	0.94	0.05	0.0	0.01
Sample3	0.17	0.83	0.0	0.0
Sample4	0.43	0.04	0.01	0.52
Sample5	0.02	0.69	0.1	0.19
Sample6	0.36	0.13	0.11	0.4

(b)

	Gene1	Gene2	Gene3	Gene4	Gene5
Cell1	0.47	0.64	0.92	0.21	0.83
Cell2	0.22	0.83	0.81	0.56	0.63
Cell3	0.88	0.95	0.11	0.46	0.69
Cell4	0.12	0.26	0.73	0.81	1.0

(c)

Figure 2.3: Example of matrices **G** (a), **P** (b) and **S** (c).

important given the increasing need to deal with large datasets. Moreover, many of these approaches cannot predict critical cell-type-specific gene expression. Because of this restriction, Scaden and other approaches are difficult to interpret.

However, one of the latest papers about deep learning for cell-type deconvolution tries to overcome these limitations by creating TAPE, using DNNs. It is based on an encoder that can learn higher-order latent representations, and a decoder can realize the interpretability of the output in the framework of the auto-encoder [3].

Having all in mind, in the following sections, most of the mentioned methods would be further explained to have a better approach to the state of the art of deconvolution methods for bulk data genetic expression. There are two methods not included here, SCDC and TAPE, because they will be explained in detail later on in the Methods Section, as they have been chosen for performing the comparison experiments during this EDP.

2.2 Statistical learning-based methods

Starting with the previously mentioned statistical learning-based methods, we can distinguish several, sorted by publication date, although most of them are from 2019 and 2020.

2.2.1 Bseq-SC

Bseq-SC is one of the first deconvolution software packages, released in 2016. Bseq-SC is a deconvolution method that identifies cell-type-specific GEPs from bRNA-seq data [12]. The

software uses a Bayesian hierarchical model to estimate the cell-type-specific GEPs and the proportion of each cell type in the bulk sample. The model incorporates prior information on the cell-type-specific GEPs obtained from scRNA-seq data, which helps to improve the accuracy of the deconvolution process by providing a basis for comparison with the bRNA-seq data. The software output includes the estimated cell-type-specific GEPs and the proportion of each cell type in the bulk sample.

It is worth noting that the model includes a term for technical noise, which is assumed to be normally distributed with a mean of zero and a variance that is estimated from the data. The model also includes a term for batch effects, which are assumed to be normally distributed with a mean of zero and a variance that is estimated from the data. The batch effect term is included to account for any systematic differences in gene expression that may be due to differences in sample preparation or sequencing conditions. The model estimates the cell-type specific expression and the proportion of each cell type in the bulk sample while accounting for technical noise and batch effects. This helps to improve the accuracy of the deconvolution process by reducing the impact of these sources of variation on the estimated GEPs and cell type proportions.

However, this model has some disadvantages, for instance, the requirement of prior knowledge of single-cell reference GEP, which may not be available for all cell types or tissues. Also, it may be unable to account for all sources of variation in the bRNA-seq data, which may affect the accuracy of the estimated cell-type specific GEPs and cell-type proportions.

2.2.2 MuSiC

The MuSiC method is a computational approach that utilizes cross-subject scRNA-seq data to estimate cell type proportions in bRNA-seq data [13]. MuSiC addresses the challenge of estimating cell type proportions in bulk samples when single-cell reference data is available from multiple individuals or subjects.

MuSiC leverages the scRNA-seq data from multiple subjects to generate a shared cell type-specific gene expression signature, which is then used to deconvolve the cell type proportions in bRNA-seq data from a target subject. The method involves several steps, which can be summarized in the overview in Figure 2.4. MuSiC creates a hierarchical clustering tree showing the similarity between cell types. Using this tree structure, the user can identify the subsequent steps of recursive estimation and determine the appropriate cell types to be grouped together at each stage. Next, MuSiC identifies the genes that exhibit consistent grouping within each

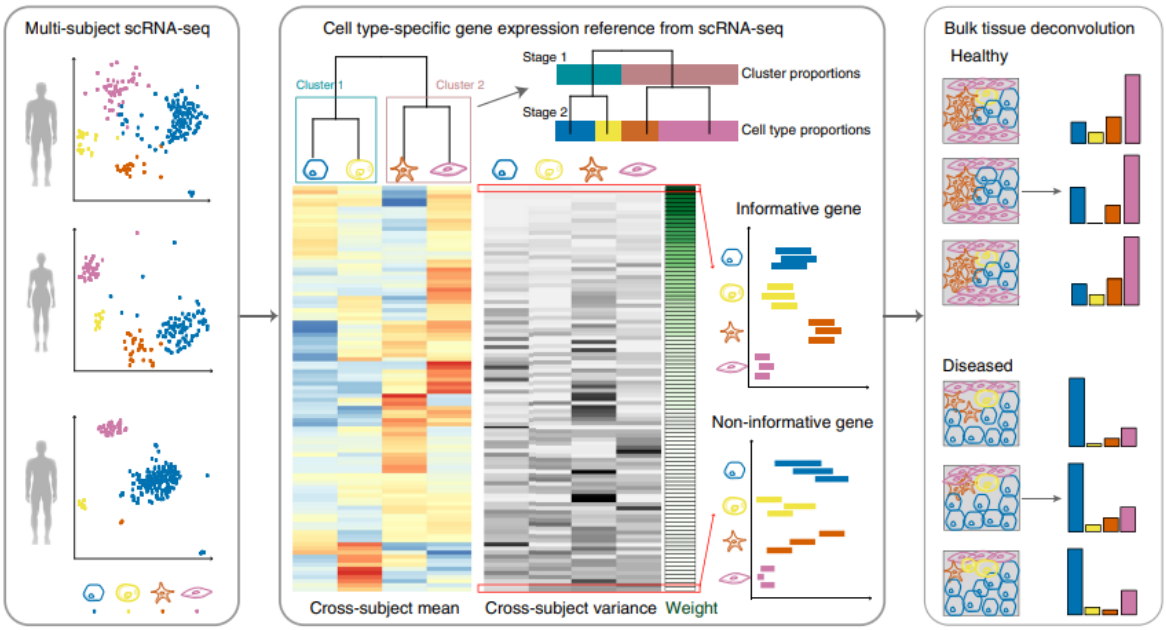


Figure 2.4: Overview of MuSiC framework (from [13], licensed by CC BY 4.0).

cluster and calculates the cross-subject mean (represented by a color gradient from red to blue) as well as the cross-subject variance (represented by a color gradient from black to white) for these genes within each cell type.

The method is based on W-NNLS that does not rely on previously chosen marker genes, contrary to CSx, which is based on NNLS but uses a reference gene expression signature matrix. Indeed, the iterative estimate technique gives more weight to informative genes (with low cross-subject variance) and less weight to non-informative genes by default. As it is a linear regression-based technique, genes with low cross-cell type changes will have limited leverage and hence have less effect on the regression, whereas genes with significant weight and leverage will have the strongest influence.

One of the key advantages of MuSiC is that it accounts for inter-individual variability in cell type proportions, which can arise due to biological differences or technical variability between individuals. By leveraging scRNA-seq data from multiple subjects, MuSiC aims to improve the accuracy of cell type proportion estimation in bRNA-seq data by accounting for inter-subject variability.

However, like any computational method, MuSiC also has some limitations. These may include the reliance on the availability of scRNA-seq data from multiple subjects, which may not always be feasible or readily available. MuSiC may also be sensitive to the quality and vari-

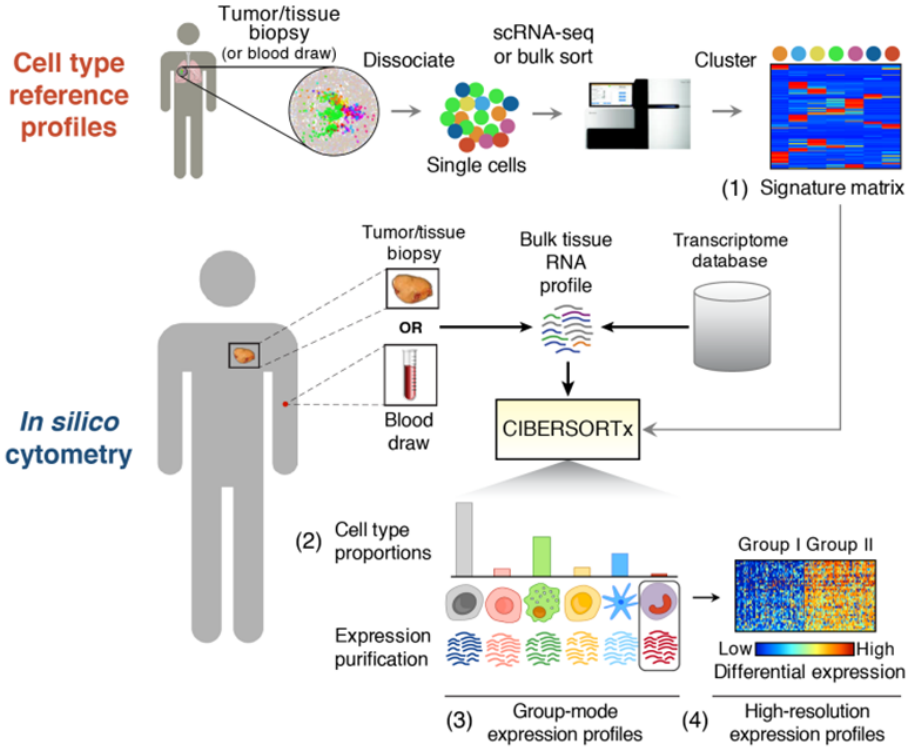


Figure 2.5: Framework for in silico cell enumeration and purification (from [14], with permission of the publisher).

ability of scRNA-seq and bRNA-seq data, and the accuracy of cell type proportion estimation can be influenced by various factors, such as batch effects, cell type-specific gene expression changes across individuals, and sample size. Additionally, MuSiC assumes that the cell type-specific gene expression signature is shared across subjects, which may not always hold true in certain biological or experimental contexts.

2.2.3 CSx

CSx is a computational method that allows the estimation of cell-type-specific GEPs from bulk tissue gene expression data [14], without the need for physical cell isolation. It is important to notice that it is an extension of the original CIBERSORT algorithm [19], a previous method developed by the same researchers for enumerating cell composition from tissue GEP.

The CSx method is designed to deconvolve bulk tissue gene expression data, such as RNA-seq or microarray data, into cell-type-specific expression profiles. It uses a reference gene expression signature matrix that represents the GEPs of different cell types. This reference

matrix is used to infer the relative abundance of different cell types in the bulk tissue sample, as well as their corresponding GEP. In Figure 2.5 the described framework can be observed.

CSx utilizes the NNLS regression algorithm, an optimization framework to solve the least squares problem with non-negativity constraints, to infer the cell-type-specific expression profiles while accounting for the mixed cell populations in the bulk tissue.

One of the strengths of CSx is its ability to infer cell-type-specific GEPs without requiring physical cell isolation, which can be challenging and time-consuming in some cases. CSx has been widely used in various fields of research, including cancer immunology, neurobiology, and immunotherapy research, to study the composition and dynamics of cell populations within complex tissues.

Nevertheless, it must be emphasised that while CSx can provide valuable insights into cell-type-specific GEP, it has limitations. It relies on the availability of an appropriate reference gene expression signature matrix, and its accuracy depends on the quality of the reference matrix and the bulk tissue gene expression data. Also, as stated in the paper: *'Although NNLS is robust on simple mixtures and toy examples, its performance on more complex mixtures inherent within real tissue samples can be affected by noise, imprecision, and missing data in the linear system'*. Then, this algorithm has limitations on sensitivity to noise and uncertainties in gene expression data, and, due to the linearity, potential for bias, since it assumes that the relationship between the gene expression data and the reference signatures is linear.

2.2.4 DWLS

DWLS method has been proposed to estimate cell-type composition from bulk gene expression data using a scRNA-seq-derived cell-type signature [15]. DWLS aims to address the challenges of estimating cell-type composition from bulk gene expression data by using a weighted least squares approach, where the gene expression data of bulk samples is modelled as a linear combination of the scRNA-seq-derived cell-type signatures, with weights representing the cell-type composition of the bulk samples.

The dampened word refers to the regularization step that is incorporated into the method to prevent overfitting and improve estimation accuracy. This regularization is achieved through a dampening factor, which is used to shrink the estimated cell-type composition towards a prior expectation. The dampening factor is determined automatically based on the data and can be adjusted by the user to control the level of regularization.

DWLS has been shown to provide accurate estimates of cell-type composition from bulk gene expression data, even in the presence of noise and uncertainties. It has been applied in various studies to infer cell-type-specific gene expression and relative intra-cell type sparing, which can be valuable in understanding cell-type-specific contributions in complex tissues and diseases.

As occurs with CSx, one of the limitations of this approach is that DWLS relies on the availability of accurate and representative scRNA-seq-derived cell-type signatures. Another limitation is the inability to capture cell-type heterogeneity and dynamics because it estimates cell-type composition based on static GEPs of individual cell types, therefore, it may not fully capture the heterogeneity and dynamics of cell types within a complex tissue or biological system. This limitation is also present on CSx.

2.2.5 CDSeq

The proposed deconvolution method, CDSeq, estimates cell proportions and GEPs from mixed RNA-Seq samples. It employs a generative model that explicitly considers how reads are generated and estimates cell proportions instead of RNA proportions [2]. The model employs multinomial random variables to capture the stochastic nature of reads and therefore inherently builds in the constraint that proportions are non-negative and sum to one on the parameters of interest. CDSeq also accommodates possibly different amounts of RNA per cell from cell types whose cells differ in size when estimating the proportion of cells of each type in the sample. In addition, instead of specifying the number of cell types a priori, CDSeq provides an algorithm that allows the data to guide the selection of the number of cell types. Finally, CDSeq proposed a quasi-unsupervised learning strategy that augments the input data (GEPs from mixed samples) with additional GEPs from pure cell lines that are anticipated to be components of the mixture. In Figure 2.6 a schematic of the model approach can be observed.

Contrary to the rest of the models described until now, CDSeq does not require a single-cell reference. It estimates cell proportions and gene GEPs from mixed RNA-Seq samples using a generative model that explicitly considers how reads are generated and estimates cell proportions instead of RNA proportions. This can be considered an advantage as you may not have always the availability of single-cell reference. Another advantage is that the model is designed to handle noise in the gene expression data because it employs multinomial random variables to capture the stochastic nature of reads.

Some of the limitations and potential disadvantages are its limited ability to handle corre-

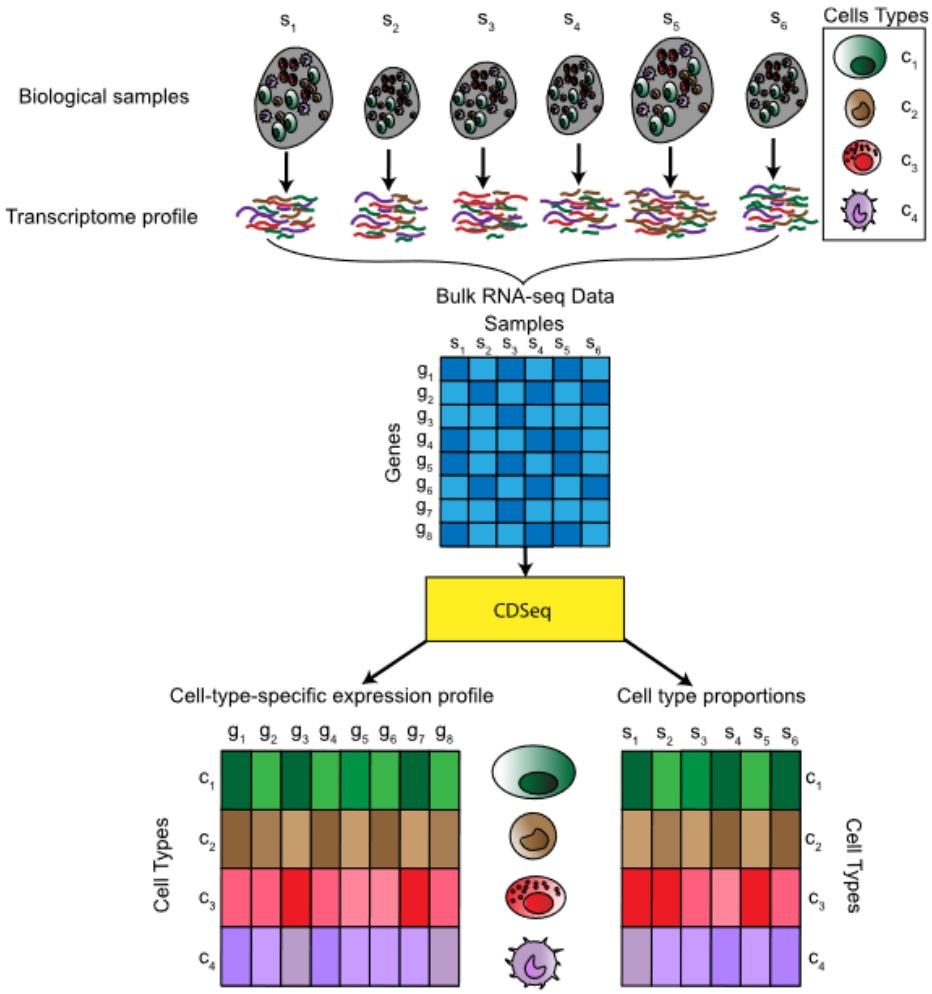


Figure 2.6: Schematic of the CDSeq approach (from [2], licensed by CCO 1.0).

lated gene expression, as CDSeq assumes that gene expression counts are independent, which may not be true in some cases where genetic pathways are regulated as units. Also, it has limited ability to handle batch effects. Furthermore, this model is computationally intensive because it uses a Gibbs sampler to iteratively assign a cell type to each read. However, the authors provide a data dilution strategy to speed up the algorithm. Finally, CDSeq assumes that cell types do not overlap, which may not be true in some cases where cell types share some gene expression patterns.

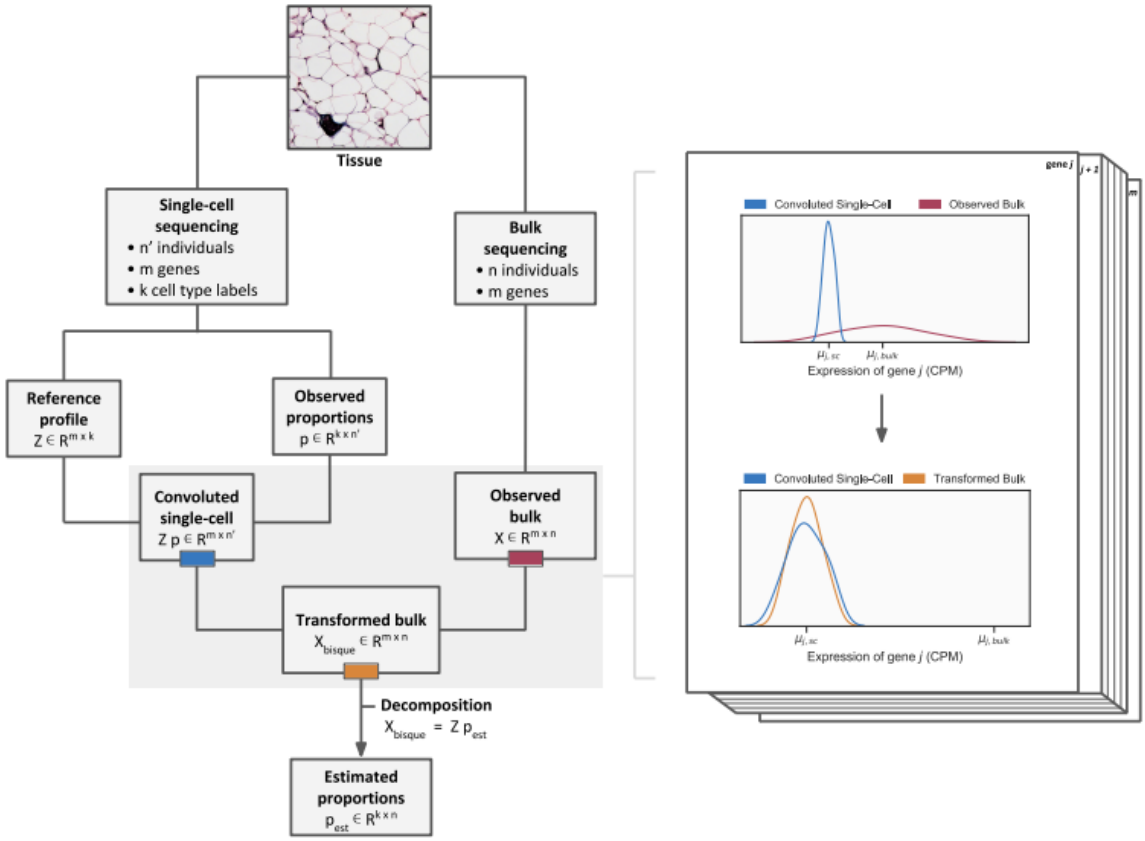


Figure 2.7: Overview of Bisque framework (from [16], licensed by CC BY 4.0).

2.2.6 Bisque

Bisque is a regression-based approach that utilizes scRNA-seq data to generate a reference expression profile and learn gene-specific bulk expression transformations to robustly decompose RNA-seq data [16].

Bisque works by first constructing a reference expression profile from scRNA-seq data, which serves as a representative expression profile for each gene in the sample. The reference expression profile is generated by calculating the median expression of each gene across all cells in the scRNA-seq data. This reference profile is then used as a basis to learn gene-specific bulk expression transformations. Next, given both the reference and the bulk data, the system learns gene-specific modifications of the data to account for technical biases between sequencing methods. Finally, it can employ NNLS regression to estimate cell proportions from the bRNA-seq data using the reference and transformed bulk expression data. These steps can be observed in the Figure 2.7, which shows on the right how the single-cell and bulk expressions

Deconvolution method	Single cell reference	Marker gene requirement	Handles batch effects	Handles noise
Bseq	✓	Optional	✓	✓
MuSiC	✓	No	✓	✓
CSx	✓	No	✓	✓
DWLS	✓	✓	No	No
CDSeq	No	No	No	✓
Bisque	✓	Optional	No	No
SCDC	✓	No	✓	No

Table 2.1: Comparison table of statistical learning-based deconvolution methods.

are integrated by learning gene-specific bulk transformations that align the two datasets for accurate decomposition.

Some of the characteristics that make this model worthwhile are that it learns gene-specific bulk expression transformations, which can account for variability in gene expression across cell types and technical differences between scRNA-seq and bRNA-seq data. Also, it can handle multiple cell types simultaneously, allowing for the estimation of the proportions of multiple cell types in bRNA-seq data.

Nevertheless, as with all the methods, it is limited by assuming a linear relationship between scRNA-seq and bRNA-seq data, which may not always hold true in all biological contexts. Furthermore, another fundamental assumption is that single-cell estimations of cell proportions properly represent the real proportions to estimate. Then, if the fraction of cell types obtained by single-cell tests differs greatly from the genuine physiological distributions, the accuracy may be reduced.

To conclude with the statistical learning-based methods section, Table 2.1 shows the comparison of some of the characteristics for all the methods described. As it can be seen, most of them need a single cell reference, but not of the marker gene. Also, it is curious to notice that old methods have tools for handling bach effects and noise, but some of the newest do not have them.

2.3 Deep-learning-based methods

Most of the previous algorithms rely on GEPs of cell type-specifically expressed genes to estimate cellular fractions using linear regression. It has been shown that the design of the GEP

is the most crucial factor in most deconvolution methods, as it results from different algorithms strongly correlated given the same GEP. Ideally, an optimal GEP should contain a set of genes that are predominantly expressed within each cell population of a complex sample. These genes should exhibit stable expression across experimental conditions, such as in health and disease, and be resilient to experimental noise and bias. However, bias is typically inherent in biomedical data and can arise from factors such as intersubject variability, species variations, different data acquisition methods, experimenters, or data types. The negative impact of bias on performance can be somewhat improved by using large heterogeneous GEP matrices. Consequently, recent advancements in cell deconvolution have relied primarily on sophisticated algorithms for data normalization and engineering optimal GEPs. While GEP-based approaches form the foundational basis of modern cell deconvolution algorithms, it has been proposed that DNNs could create optimal features for cell deconvolution without relying on the complex generation of GEPs [18].

DNNs are a class of machine learning models capable of learning and extracting intricate patterns and relationships from complex data. At the core of a neural network are artificial neurons, which receive input signals, apply mathematical transformations to them, and produce output signals. The strength of neural networks lies in their ability to learn from data through a process called training. During training, the network adjusts the weights and biases associated with each neuron based on the input data and the desired output. This adjustment is performed using optimization algorithms, such as backpropagation, which calculate the gradients and update the parameters to minimize the difference between the network predictions and the true values [20].

The depth of a neural network refers to the number of hidden layers it possesses. Deep neural networks, or deep learning models, typically have multiple hidden layers, allowing them to learn hierarchical representations of the input data. This hierarchical representation enables the network to capture intricate and abstract features, leading to improved performance in complex tasks, as the deconvolution of bulk data. Therefore, since 2020 the deconvolution algorithms have taken a machine learning-driven path, which may be compared favourably to existing deconvolution algorithms in both prediction robustness and accuracy [20].

2.3.1 Scaden

Scaden is a deep neural network algorithm that is trained on scRNA-seq data to accurately determine the cellular composition of tissues using gene expression information [18]. The

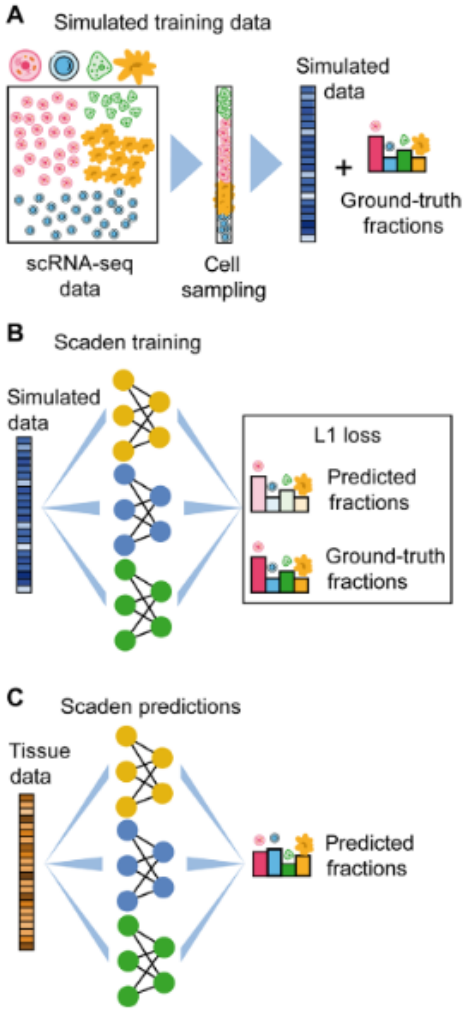


Figure 2.8: Overview of training data generation and cell type deconvolution with Scaden (from [18], licensed by CC BY-NC 4.0).

algorithm uses simulated bRNA-seq data generated from scRNA-seq data to predict cell type fractions in bRNA-seq samples. To obtain simulated data, Scaden uses scRNA-seq data to create in silico bRNA-seq data of a predefined type (target tissue) with known composition, across datasets. This is done by simulating bRNA-seq data from scRNA-seq data using a statistical model that accounts for technical noise and biological variability. The simulated data is then used to train the Scaden model, as can be observed in Figure 2.8.

The Scaden model is an ensemble of three deep neural networks with varying architectures and degrees of dropout regularization. Dropout randomly drops out a fraction of the neurons in the network during training, which forces the remaining neurons to learn more robust features that are less sensitive to noise in the data. The models are trained independently for 5000 steps

using supervised learning on datasets of simulated bRNA-seq samples. Model selection aims to find architecture and hyper-parameters that robustly deconvolve simulated tissue RNA-seq data and real bRNA-seq data. To perform deconvolution with Scaden, a bRNA-seq sample is fed into a trained Scaden ensemble, and the trained DNNs generate three independent predictions for the cell type fractions of this sample. These three predictions are then averaged per cell type to yield the final cell type composition for the input bRNA-seq sample.

The model selection process is optimized on simulated peripheral blood mononuclear cells (PBMCs) datasets to capture inter-experimental variation and simulate batch effects between datasets. Scaden uses leave-one-dataset-out cross-validation for model optimization, where a model is trained on simulated data from all but one dataset, and performance is tested on simulated samples from the left-out dataset. This allows Scaden to simulate batch effects between datasets and helps to test the generalizability of the model. However, it is important to note that Scaden is not specifically designed to handle batch effects, and its ability for it may depend on the extent and nature of the batch effects present in the data.

Nevertheless, Scaden is designed to handle noise in the data. The algorithm uses deep neural networks trained on simulated bRNA-seq data generated from scRNA-seq data to predict cell-type fractions in bRNA-seq samples. The simulated data is generated to account for technical noise and biological variability, which are common sources of noise in gene expression data. In addition, Scaden uses a regularization technique called dropout to prevent overfitting and to improve the ability of the model to handle noise. Overall, Scaden is designed to handle noise in the data and has been shown to outperform existing deconvolution algorithms in both precision and robustness.

Another strength of the model is the ability to deconvolve across data types, as it can seamlessly deconvolve microarray data of the same tissue, which is noteworthy as microarray data are known to have a reduced dynamic range and several hybridization-based biases compared to RNA-seq data. Furthermore, once trained, the prediction runtime scales linearly with sample numbers and is usually in the order of seconds, making Scaden a useful tool if deconvolution is to be performed on very large datasets.

Continuing with some limitations, although the novelty of this model is that it does not rely on GEPs, it still needs single-cell data reference for the creation of the simulated data, so, until here, there is still a challenge for deconvolution methods to only work from bulk data to single cell. Moreover, it is necessary to train a new model for deconvolution if no perfect overlap in the feature space exists. This constraint limits the usefulness of pre-trained models. Finally, the most prominent and obvious issue of Scaden is the difference between simulated scRNA-

seq data used for training and the bRNA-seq data subject to inference. While Scaden is able to transfer the learned deconvolution between the two data types and achieves state-of-the-art performance, efforts to improve this translatability could improve Scaden prediction accuracy even further.

Notwithstanding the foregoing, Scaden outperforms existing algorithms on simulated and real tissue datasets, including postmortem human brain tissue, and is robust to training data bias and species differences. Scaden also allows for the inclusion and mixing of different scRNA-seq experiments in the training dataset, further increasing its robustness. The algorithm has potential applications in understanding disease mechanisms and developing targeted therapies.

In summary, deep-learning-based methods are a groundbreaking approach that is gathering momentum nowadays, not only in this field. The algorithms described have clear advantages, however, their performance must be properly evaluated before falling into the false myth that deep learning can solve everything.

To conclude, all of these descriptions of the different algorithms have given a broad view of what the topic is about, understanding the problem of deconvolution and the several approaches that can it take. Henceforth, we will be able to explain the most current methods and their novelties, to determine if they are really worthwhile.

Chapter 3

Data and Methods

In this chapter, the methods and data description are collected. First of all, an overview of the origin and content of the datasets used is given. Next, the two algorithms tested in this EDP are exposed in detail, including the basic and fundamental equations, as well as the description of the libraries to be used.

3.1 Description of the data

Regarding the data, there are two databases needed for the accomplishment of our experiments, namely, pseudo bulk data and real bulk data. Additionally, for the normalization of data types on the TAPE method, a third element is needed. A gene length file that has the information of gene name, transcript start (base pair) and transcript end (base pair), obtained from the GitHub repository of [3].

3.1.1 Pseudo bulk data

The pseudo bulk data was obtained from the data used in the pseudobulk test from TAPE [3]. The authors rely on a single-cell dataset of mouse atlas from *Tabular Muris* [21], which consists of 20 organs and tissues with cell-type labels. The pseudo-bulk test was performed using data from three tissues (Limb Muscle, Marrow, and Lung). Additional data was not chosen since the shared cell types throughout procedures are extremely restricted (less than four cell types), making it impossible to mimic a real-life scenario. Specifically, "Limb Muscle" has 6 cell types, "Marrow" has 7 cell types, and "Lung" has 9 cell types. Furthermore, as the cell-type labels

are given for this dataset, also the label matrices for the pseudo bulk data was generated for evaluation purposes.

Pseudo-bulk expression data are defined as the sum of single-cell expression data from a selection of cells. As a result, in order to obtain pseudo-bulk data, cells need to be sampled with a predetermined cell type percentage (ground truth) and total cell number, similar to stratified sampling. When users have some prior knowledge about cell-type fractions in a certain tissue, the authors created cell-type fractions using the Dirichlet distribution. If they have no previous information, each prior weight will be the same (normal samples). Additionally, half of the produced samples corresponding to cell-type fractions contain zeros (sparse samples), because training with both normal and sparse data improves deconvolution performance. Next, the overall cell number is then multiplied by the cell fractions obtained for each sample to obtain the precise sampling number for each cell type. After, using the provided number of cells, a stratified sampling approach is employed to sample cells of each cell type. Lastly, the pseudo-bulk expression profile is generated by adding the expression values of each randomly selected single-cell expression profile of the sample [3].

Once the data is generated, it is critical to consider the data shift across different sequencing algorithms to ensure that the approaches work optimally. Therefore, all pseudo bulk datasets are generated using two separate sequencing protocols, namely, 10X-seq (UMI-based method) and Smart-seq (counts-based method), to perform further tests in both cross-protocol and same-protocol deconvolution. To maintain consistency, the authors processed all representations into gene names through BioMart [22].

3.1.2 Real bulk data

With respect to real data, these were the datasets used: scRNA-seq data of human pancreatic islets from Baron et al. [12] with GEO accession number GSE84133 and from Segerstolpe et al. [23] with accession number E-MTAB-5061; bulk RNA-seq of human pancreatic islet from Fadista et al. [24] with GEO accession number GSE50244. As they were all obtained from the experiments performed on the SCDC article, following the guideline in [25], and the data was firstly downloaded with raw read-count metrics. Therefore, its use is possible on both algorithms without further processing.

3.2 Methods

The methodology followed for the performance of the experiments was mainly based on the implementation of the two algorithms to be compared, SCDC and TAPE, and a later calculation of validation metrics from the results obtained. The reason for choosing these algorithms is because they are the lastest state-of-the-art methods for each of their categories, and therefore, it is assumed that they are the most updated and innovative algorithms. Some steps were followed for the implementation and use of each of them, described in the following sections. Finally, the metrics for the validation chosen are explained.

3.2.1 Method 1: SCDC

The SCDC method is a deconvolution approach that utilizes scRNA-seq data from multiple references to estimate cell-type proportions in bRNA-seq data. SCDC primarily adheres to the W-NNLS framework proposed by MuSiC, although it incorporates several distinct features concerning constructing the basis matrix [17].

To understand the W-NNLS framework on which this method is based, it is necessary to return to the root algorithm, the NNLS algorithm. It is a numerical optimization method used to solve linear regression problems where the coefficients are constrained to be non-negative. The algorithm works by iteratively solving a sequence of unconstrained least squares problems subject to the non-negativity constraint. That is, given a matrix $\mathbf{A}$ and a (column) vector of response variables $\mathbf{y}$, the objective is to find

$$\arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \text{ subject to } \mathbf{x} \geq \mathbf{0} \tag{3.1}$$

where $\|\cdot\|_2$ denotes the Euclidean norm [26].

From here, W-NNLS is a modification of NNLS that incorporates weights to adjust the contribution of each subject to the construction of a basis matrix. Specifically, the first distinct feature from MuSiC basis matrix construction relies here upon when W-NNLS calculates the maximal variance weight (MVW) per gene, which reflects the data quality, and scales the raw single-cell read-count matrix so that residuals from genes with larger weights have a more negligible impact on cell-type composition estimation. Extensively, SCDC first estimates $\hat{\sigma}^2_{gkd}$, which is a term that captures the cross-cell variation for gene g of cell type k within individual

d. Within-subject variance for subject d is then calculated as

$$\sigma^{*2}_{gd} = \max_{k}(\hat{\sigma}^{2}_{gkd}) \tag{3.2}$$

and the MVW Δ_{gd} is given by:

$$\Delta_{gd} = \frac{\sigma^{*2}_{gd}}{\underset{g'}{\text{median}}(\sigma^{*2}_{g'd})} \tag{3.3}$$

This adjustment ensures that residuals from genes with higher weights have a lesser impact on the estimation of cell-type composition, which improves the accuracy of cell-type proportion estimation by reducing the impact of noisy or low-quality data. Continuing with the differences from MuSiC, SCDC does not presume cell-type memberships as given, instead, it employs an initial SCDC run to identify and eliminate potentially misclassified cells and doublets, thereby enhancing its robustness. And finally, SCDC accommodates single-subject scRNA-seq input, wherein the direct estimation of cross-subject variance is not possible [17].

As established in the introduction, SCDC also deconvolutes the bulk gene expression data as a matrix decomposition problem, so, the gene expression levels can be found by solving the Equation 2.1 through the W-NNLS approach. However, during this algorithm, the matrices get transformed, which does not affect the final result, just it must consider the position of the rows and columns. So the final equation for the SCDC approach is:

$$\mathbf{G}^{T} = \mathbf{P} \cdot \mathbf{S}^{T} \tag{3.4}$$

Additionally, what makes SCDC stand out is the ENSEMBLE approach (Figure 3.1). When multiple scRNA-seq reference sets are accessible, by assigning greater weights to the scRNA-seq data that exhibit closer resemblance to the bRNA-seq data, SCDC consolidates deconvolution outcomes across datasets, implicitly tackling the issue of batch-effect interference. Therefore, by optimizing the predicted gene expression levels, ENSEMBLE integrates all deconvolution results [17] as:

$$(\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_{\mathrm{R}}) = \underset{(w_1, w_2, \ldots, w_R)}{\text{argmin}} \left\| \mathbf{G} - w_1\hat{\mathbf{G}}_1 - w_2\hat{\mathbf{G}}_2 - \ldots - w_{\mathrm{R}}\hat{\mathbf{G}}_{\mathrm{R}} \right\|_1 \tag{3.5}$$

The SCDC method has been demonstrated to perform well in handling complex and diverse cell types, making it suitable for various biological contexts where accurate estimation of cell-type proportions from bRNA-seq data is challenging. However, it also has limitations.

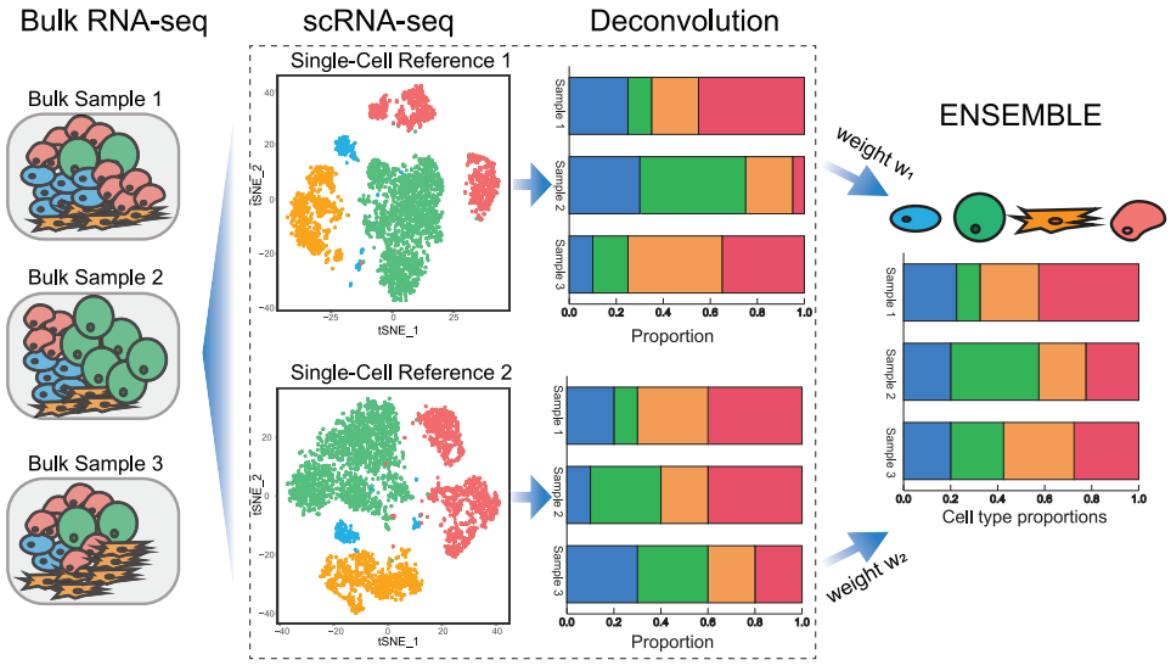


Figure 3.1: Overview of deconvolution via ENSEMBLE by SCDC (from [17], licensed by CC BY-NC 4.0).

It requires scRNA-seq data from multiple references, assumes a linear relationship between bRNA-seq and scRNA-seq data, and may be influenced by data quality and experimental design. These limitations should be considered when interpreting the results obtained from SCDC or any other deconvolution method.

Continuing with its implementation, the SCDC algorithm is collected as an R module as described on [25]. The main function for bRNA-seq data deconvolution receives three important inputs: bRNA-seq data, scRNA-seq and a list of names of the cell types. For both bRNA-seq data and scRNA-seq data, SCDC takes *ExpressionSet* objects with raw read counts as input. An *ExpressionSet* object is based on the package *Biobase*, which mainly stores data matrix (genes by samples), *featureData* (information about genes), and *phenoData* (information about samples, like the clustering results, subjects, and conditions). Although the module SCDC also has a function for generating pseudo bulk data, the data used was the one already generated from the TAPE article.

Therefore, regarding the pseudo bulk experiment, all data counts were processed to generate adequate *ExpressionSet* objects, which were then introduced into the main deconvolution function joint with the list of cell types, which were obtained from the label matrix corresponding to each pseudo bulk data. Concretely, for Limb Muscle tissue, the cells evaluated are the

skeletal muscle satellite cell, mesenchymal stem cell, endothelial cell, B cell, macrophage, and T cell; for Lung tissue cell types they are the lung endothelial cell, stromal cell, classical monocyte, myeloid cell, B cell, T cell, natural killer cell, leukocyte, and ciliated columnar cell of the tracheobronchial tree; and for Marrow tissue, we have the hematopoietic precursor cell, granulocyte, immature B cell, late pro-B cell, monocyte, macrophage, and basophil.

As mentioned, SCDC takes *ExpressionSet* objects with raw read counts, which were present in the pseudo bulk data obtained by the Smart-Seq protocol (count-based method). However, UMI-based data was also tested in order to evaluate cross-protocol performance.

Concerning the real data experiment, bRNA-seq data and scRNA-seq datasets were directly obtained as *ExpressionSet* objects, so further preprocessing was not needed. Next, data is introduced on the function jointly with the list of cell types, in this case, as we are talking of pancreatic islets datasets, the cell types were alpha, beta, delta, gamma, acinar, and ductal. Finally, the outputs of the SCDC deconvolution function are the proportion estimation matrix, the basis matrix, and the GEP of bulk data.

3.2.2 Method 2: TAPE

TAPE is a deep-learning algorithm for digital tissue dissection that aims to accurately and sensitively deconvolve bulk gene expression data into cell-type-specific gene expression profiles. Its innovative approach based on encoder-decoder architecture enables an interpretable result, contrary to many other deep learning-based solutions. Also, it introduces a new training scheme in the adaptive stage, where the model is trained on real bulk data. This architecture makes TAPE highly accurate and sensitive in capturing biologically significant changes in clinical data, and it allows for the identification of potential gene expression differences at the cell-type level [3].

The use of TAPE needs three stages. The first one involves generating pseudo-bulk data from a single-cell dataset to train the model. The goal is to create a dataset that represents bulk gene expression data, which is the sum of single-cell expression data from a subset of cells.

To generate the pseudo-bulk data, the authors use single-cell expression data with cell type fractions. They first generate cell-type fractions using the Dirichlet distribution, which is a probability distribution over the simplex of non-negative real numbers that sum to one. The Dirichlet distribution allows users to define the prior cell fractions by setting the parameters in the Dirichlet function. If users do not have prior knowledge, the prior weight of each cell type will be the same (normal samples).

Next, the authors multiply the total cell number with the generated cell fractions for each sample to acquire the exact sampling number for each cell type. After that, they use a stratified sampling method to sample cells of each cell type with the given number. Finally, the pseudo-bulk expression profile is created by summing the expression values of the randomly selected single-cell expression profiles for each sample.

The second stage is the training stage, where authors try to train the model to return the adequate cell fractions after the encoder, for later use of the cell fractions to reconstruct the bulk profile. Following the problem described in the introduction and Equation 2.1, the architecture of TAPE can be defined as:

$$\begin{aligned} f_\phi(\mathbf{G}) &= \tilde{\mathbf{P}} \\ f_\psi(\tilde{\mathbf{P}}) &= \tilde{\mathbf{P}} \cdot \mathbf{S} \\ f_\psi\left(f_\phi(\mathbf{G})\right) &= \tilde{\mathbf{G}} \end{aligned} \tag{3.6}$$

where f_ϕ and f_ψ are two coordinated deep neural networks, and the tilde on symbols represents the outputs. The encoder (f_ϕ) takes the bulk gene expression data as input and compresses it into a lower-dimensional representation, which is a set of cell-type proportions being trained to predict the cell-type ratios accurately. The decoder (f_ψ) takes the cell-type proportions as input and reconstructs the bulk gene expression data. It is a natural cell-type-specific signature matrix that can be learned after the training stage and then adapted to the bulk data after the adaptive stage.

Finally, the last stage is the adaptive stage, where the model is fine-tuned on new data using an unsupervised training scheme. In this stage, the model is required to predict the cell-type-specific gene expression profiles and cell-type proportions on real bulk data. To achieve this, TAPE uses a greedily iterative optimizing method. Having the loss functions defined as:

$$\begin{aligned} \mathrm{MAE}(\mathbf{P}, \tilde{\mathbf{P}}) &= \frac{\sum_{ij} \left|\mathbf{P}_{ij} - \tilde{\mathbf{P}}_{ij}\right|}{n \times k} \\ \mathrm{MAE}(\mathbf{G}, \tilde{\mathbf{G}}) &= \frac{\sum_{ij} \left|\mathbf{G}_{ij} - \tilde{\mathbf{G}}_{ij}\right|}{n \times k} \end{aligned} \tag{3.7}$$

In the first step, the decoder is optimized with the mean absolute error (MAE) loss function between the ground truth bulk gene expression data and the predicted bulk gene expression data, as well as the MAE loss function between the predicted cell-type-specific gene expression profiles and the cell-type-specific gene expression profiles learned during the training stage.

In the second step, the encoder is optimized with the MAE loss function between the ground

input : Encoder parameters E and decoder parameters D from the initial training stage, GEPs of bulk RNA-seq $\mathbf{B}$ of size $n \times m$, step number α, max iteration β
output: signature matrix $\mathbf{S}$ of size $k \times m$, predicted fractions $\mathbf{X}$ of size $n \times k$, training loss L

$\tilde{\mathbf{S}}_0, \tilde{\mathbf{X}}_0 \leftarrow \texttt{model}(\mathbf{B})$;
for $k \leftarrow 1$ **to** β **do**
  for $i \leftarrow 1$ **to** α **do**
    $\tilde{\mathbf{B}}, \mathbf{X} \leftarrow \texttt{model}(\mathbf{B})$;
    $L \leftarrow \texttt{MAE}(\tilde{\mathbf{B}}, \mathbf{B}) + \texttt{MAE}(\mathbf{S}, \tilde{\mathbf{S}}_0)$;
    $D \leftarrow D - \frac{\partial L}{\partial D}$;
  end
  for $j \leftarrow 1$ **to** α **do**
    $\tilde{\mathbf{B}}, \mathbf{S} \leftarrow \texttt{model}(\mathbf{B})$;
    $L \leftarrow \texttt{MAE}(\tilde{\mathbf{B}}, \mathbf{B}) + \texttt{MAE}(\mathbf{X}, \tilde{\mathbf{X}}_0)$;
    $E \leftarrow E - \frac{\partial L}{\partial E}$;
  end
end
$\mathbf{S}, \mathbf{X} \leftarrow \texttt{model}(\mathbf{B})$;

Figure 3.2: TAPE adaptive training procedure (from [3], licensed by CC BY 4.0).

truth cell-type proportions and the predicted cell-type proportions. The optimization continues until the MAE loss function between the ground truth cell-type proportions and the predicted cell-type proportions does not decrease. This procedure can be better visualized in Figure 3.2, where there is a representation of the algorithm operation.

Regarding its implementation and use, TAPE can be installed as a Python library as explained on [27]. In contrast to SCDC, its use requires more than three inputs:

- scRNA-seq data (cell types should be labeled) in either TPM (recommended), FPKM or counts measurement. The indices are cell types, the columns are gene names.
- bRNA-seq data with indices of sample name and columns of gene names.
- Gene length file: TAPE has the ability to normalize data into TPM/FPKM format thanks to this data.
- Datatype: 'TPM', 'FPKM' or 'counts'. The parameter specifies the method to normalize pseudo-bulk data for training.
- Mode: 'overall' or 'high-resolution'. The returned signature matrix in the 'high-resolution'

mode will be a dictionary-like 'cell type': gene expression data,..., and the gene expression data is a data frame. In the 'overall' mode, the signature matrix is a data frame with cell types as indices and gene names as columns. Set to 'overall' for comparison with the other method.

- Adaptive: True or False. If adaptive is False, it will not estimate the signature matrix and will return None. If adaptable is set to True, the mode determines the format of the returned signature matrix. Set to True during the experiments because we need the signature matrix for calculation of the GEP.
- Batch size: integer related to training result.
- Epochs: integer related to training result.

Regarding our pseudo bulk experiment, TAPE deconvolution function was used with the following parameters: datatype 'counts' for either UMI method or counts method, mode overall, adaptive true, batch size and epochs 128 (default value). With respect to our real data experiment, after adapting the *Expressionset* data to data frames (adequate for this function), the function was used with the same parameters.

Finally, the model outputs two data frames representing the proportion estimation matrix and the basis matrix. Contrary to SCDC, the GEP is not calculated implicitly on the model but can be easily get by solving the matrix problem of Equation 2.1.

3.2.3 Error and correlation validation

In order to evaluate the performance of the algorithms, it is difficult to measure it fairly in all cases using only one metric, so both error and correlation are calculated. For instance, assume that there are only two cell types in the tissue and that the percentage of one type in ground truth ranges from 80-90%. If the model forecasts that this cell type proportion is 100%, the correlation value may suggest satisfactory performance, but the error value may indicate the reverse. To prevent the problem of rejecting fractions of minor cell types, both metrics are combined [3].

On the one hand, regarding correlation values, we can find several metrics. The CCC score measures the agreement between two continuous variables. It takes into account both the correlation (strength) and the bias (deviation from the line of perfect agreement) between the two variables. It is often used in inter-rater reliability studies or when comparing two measurement

methods. The CCC score ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement, and -1 indicates perfect disagreement [28]

The Pearson correlation coefficient measures the linear relationship between two continuous variables. It assumes that the relationship between the variables is linear and that the data follows a bivariate normal distribution. It quantifies both the strength and the direction of the linear relationship between the variables. The Pearson correlation coefficient ranges from -1 to 1 [29].

Finally, Spearman's correlation coefficient assesses the monotonic relationship between two variables, which means it measures the consistency of the rank order of the data. It calculates the correlation based on the ranks of the data rather than the actual values. Spearman's correlation coefficient also ranges from -1 to 1, with similar interpretations to the Pearson correlation coefficient [29].

Having all in mind, the most accurate metric for the matter of hand is the CCC as it measures the concordance between the predicted fraction and the ground truth, not assuming linear relationships and using only the real values.

On the other hand, we can find several error metrics. The MAE calculates the average absolute difference between the predicted proportions and the true proportions for each cell type. It provides a measure of the overall magnitude of the errors. Mean Squared Error (MSE) calculates the average squared difference between the predicted proportions and the true proportions for each cell type. It gives more weight to larger errors and is useful when you want to penalize larger deviations. Finally, the Root Mean Squared Error (RMSE) is the square root of the MSE and provides a measure of the average magnitude of the errors, similar to the MAE but giving more weight to larger errors. for the sake of our experiments, the most adequate metric is the MAE as measures the absolute errors, not penalising deviations, which in the case of cell types proportions could be a normal scenario.

To conclude, the combination of CCC and MAE metrics could result in a suitable scenario for the evaluation of our methods, so we define MAE as on Equation 3.7 and CCC as:

$$\mathrm{CCC}(p,\tilde{p}) = \frac{2 \times \mathrm{cov}(p,\tilde{p})}{\sigma_p^2 + \sigma_{\tilde{p}}^2 + (\mu_p - \mu_{\tilde{p}})} \tag{3.8}$$

where $\mathrm{cov}(p,\tilde{p})$ stands for the covariance between these two vectors. These two measures are applied to all predicted matrix $\tilde{\mathbf{P}}$ and ground truth matrix $\mathbf{P}$ data points. To be more precise, we restructure the matrix into a vector and then calculate the total CCC between two vectors. This

style of computation frequently yields a greater CCC value than computing the average CCC value for each cell type [3].

3.2.4 Bland-Altman plot

The Bland-Altman plot, also known as a Tukey mean-difference plot, is a graphical method used to assess the agreement or disagreement between two measurement methods or assays. In this case, it will be used to assess the agreement between predicted and ground truth proportions. So, it provides a visual representation of the differences between the two types of data by plotting the average of the measurements, or the ground truth values, on the x-axis and the differences between the measurements on the y-axis. The plot typically consists of a scatter plot of the data points, with the mean difference and limits of agreement displayed as reference lines.

To construct a Bland-Altman plot, the first step is to gather a set of paired measurements from the two methods or instruments being compared. In this case, ground truth proportions of cell types and predicted proportions by each deconvolution method. Then, for each pair of measurements, the difference is calculated. Finally, a scatter plot is created with the ground truth values on the x-axis and the differences between the measurements on the y-axis. It is essential for the comprehension of the plot to add reference lines, representing the differences between the two measures and the limits of agreement, which are typically set as the mean difference ± 1.96 times the standard deviation of the differences. These lines help identify the range within which most of the differences between the measures lie.

By examining the plot, any systematic bias can be identified, like consistent over- or under-estimation by one method, the presence of outliers, or heteroscedasticity, which refers to the unequal spread of differences across the range of measurements [30].

3.2.5 Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test where the average of two dependent samples is statistically analyzed to see whether there are any significant differences. It is a paired difference test, similar to the paired Student's t-test, but it is used when the data distribution is not normal, a condition to be met on T-test.

The hypotheses established for the Wilcoxon signed-rank test for paired data are:

- Null hypothesis (H0): the difference between the paired observations in the population is

zero. In our case, the absolute errors on method TAPE are equal to the absolute errors on method SCDC.

- Alternative hypothesis (H1): the difference between the paired observations is not equal to zero, that is, the absolute errors on method TAPE are not equal to the absolute errors on method SCDC.

This type of test is used to compare the final performance of both methods on the pseudo bulk experiments, as is where the ground truth data is known. Then, as the data for the Wilcoxon test consists of a set of paired observations, in our case, the first pair consists of a vector created from the concatenation of absolute errors from method TAPE, for each protocol scenario, and the second pair consists of another vector from the concatenation of absolute errors but from SCDC.

As a summary of how the test is calculated, first, the differences between the paired observations are ranked in absolute terms. Then, the positive and negative ranks are summed separately. Finally, the test statistic is calculated as the smaller of the two sums. Considering $\alpha = 0.05$, if the calculated p value is smaller than 0.05, we would have sufficient evidence to reject the null hypothesis and accept the alternative, having the conclusion that absolute errors on method TAPE are not equal to the absolute errors on method SCDC [31].

Chapter 4

Results

The findings of the many experiments undertaken in the present EDP are provided in this section. Firstly, in the experiment with pseudo bulk data, the deconvolution results of each tissue are examined independently for both approaches, and a final comparison of both methods is provided. Continuing with the real-data experiment, the results are analyzed independently having in mind the conclusions taken from the previous experiment, and a final comparison is described given these results.

4.1 Experiment with pseudo bulk data

As stated in Section 3.1.1, pseudo bulk data is a known solution dataset built to verify the performance of the methods by comparing with ground truth data. Then, the objective of this experiment is to try all different data from different tissues on both algorithms and, by calculating MAE and CCC and observing Bland-Altman plots, to compare the performance of the two methods. Furthermore, library preparation techniques differ amongst laboratories, and bulk deconvolution studies are frequently carried out utilizing single-cell reference supplied by others. So, to have a more realistic scenario, a cross-protocol experiment is also performed for all the tissues and algorithms, having a more global vision of how the algorithms operate.

4.1.1 SCDC pseudo bulk experiment

For the statistical learning-based method of choice, the first step in the pseudo bulk experiment was processing the data. Raw read counts on txt format were transformed into *ExpressionSet*

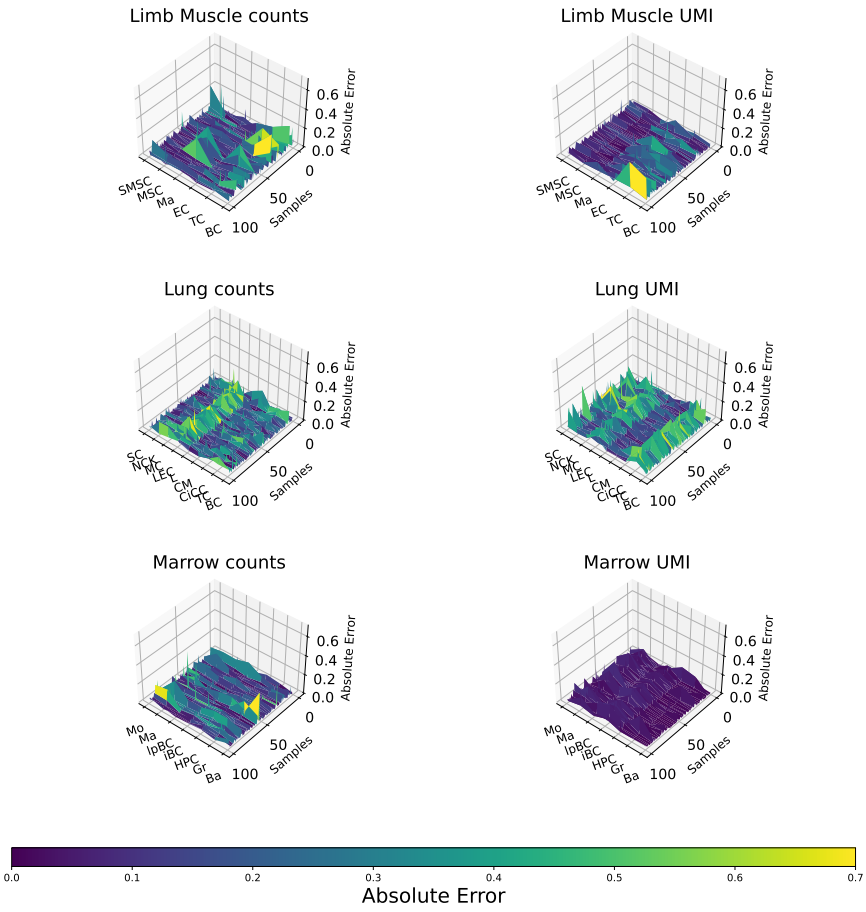


Figure 4.1: Absolute error for SCDC **P** matrix and same protocol scenario.

objects for both bRNA-seq data and scRNA-seq data. Then, the cell types list was obtained from the known solution label matrix dataset. Finally, the experiment concludes with the deconvolution function, which outputs the needed files, saved as comma-separated values (CSV) files for later evaluation. These steps are repeated on the three types of tissues and changing the inputs of the function to conduct the same protocol and cross-protocol experiment.

Continuing with the evaluation of the experiment, we focus on evaluating the accuracy and reliability of predicted data of proportions of cell types (matrix **P**) by comparing it with the available ground truth of the matrix of labels. An initial visualization in 3D of the absolute error for the predicted proportions on each scenario is shown in Figures 4.1 and 4.2. Each figure

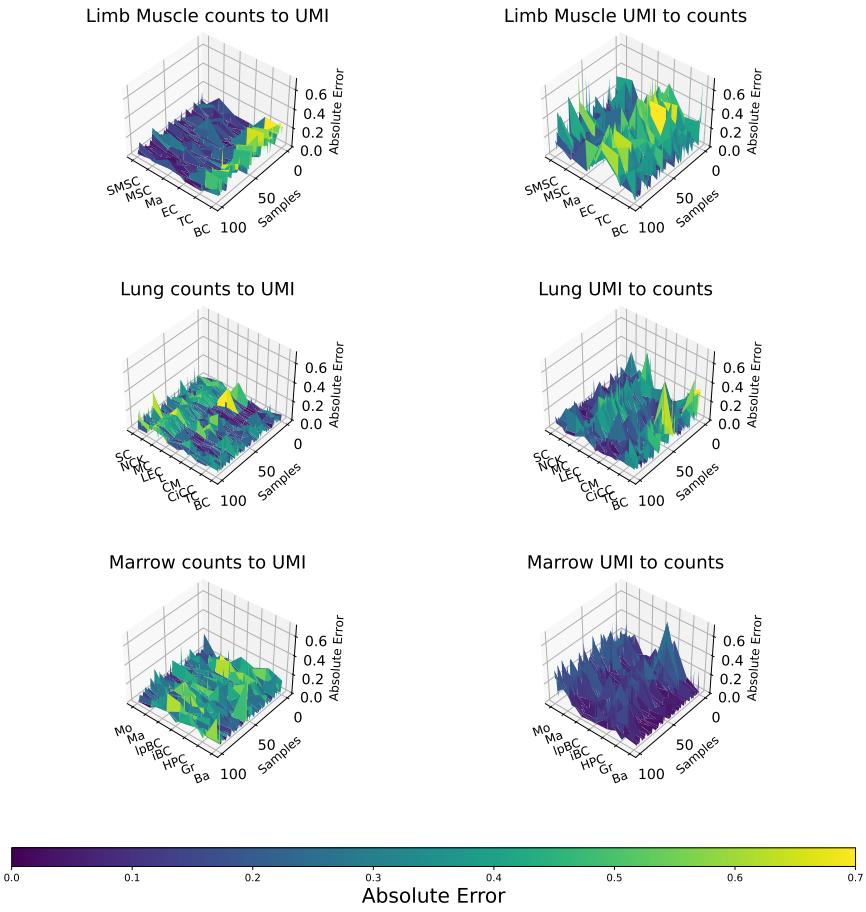


Figure 4.2: Absolute error for SCDC **P** matrix and cross-protocol scenario.

represents six subplots for each of the tissues and protocols tested in the experiments. X-axis is representing the cell types, being acronyms following this rule: B cell (BC), T cell (TC), endothelial cell (EC), macrophage (Ma), mesenchymal stem cell (MSC), skeletal muscle satellite cell (SMSC), ciliated columnar cell of tracheobronchial tree (CiCC), classical monocyte (CM), leukocyte (L), lung endothelial cell (LEC), myeloid cell (MC), basophil (Ba), granulocyte (Gr), hematopoietic precursor cell (HPC), immature B cell (iBC), late pro-B cell (lpBC), and monocyte (Mo).

Regarding the first figure, it refers to the same-protocol scenario, and it can be observed that most of the absolute error values lie from 0 to 0.4, with low peaks on the plots. However, the

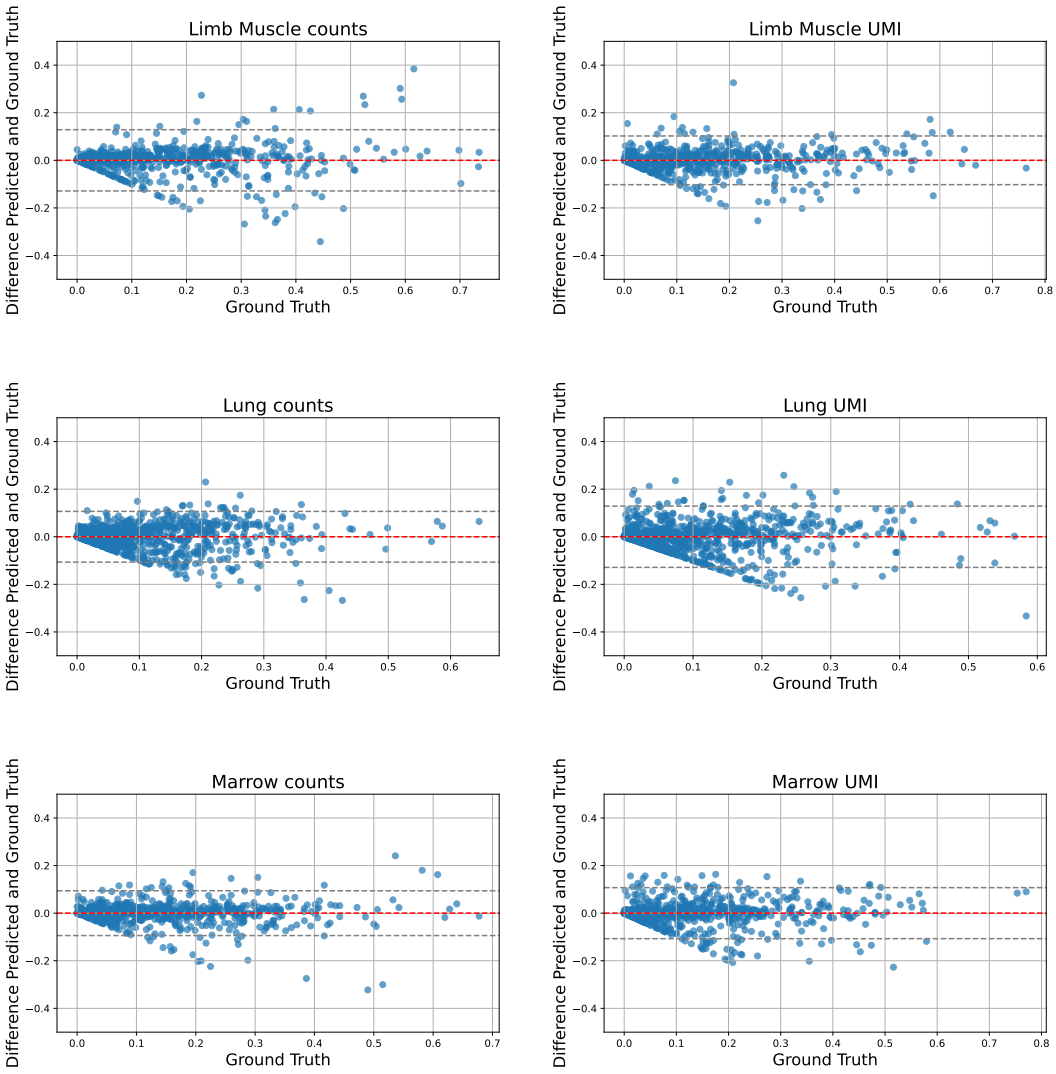


Figure 4.3: Bland-Altman plots for SCDC approach and same-protocol scenario

second figure represents the cross-protocol scenario, where more yellow (near 0.7 error) peaks are observed. Between tissues, the Marrow tissue has, in general, less absolute error than the other two. In summary, there is a clear difference between the absolute error of cross-protocol and the same-protocol scenario, being the cross-protocol much higher.

After evaluating the absolute error, the Bland-Altman plots show the level of agreement between predicted and ground truth proportions. By observing this kind of plot, the heteroscedasticity or homoscedasticity of the samples can be deduced. A model presents heteroscedasticity when the variance of the errors is not constant in all the observations made, which usually implies non-compliance with one of the basic hypotheses on which the linear regression model is

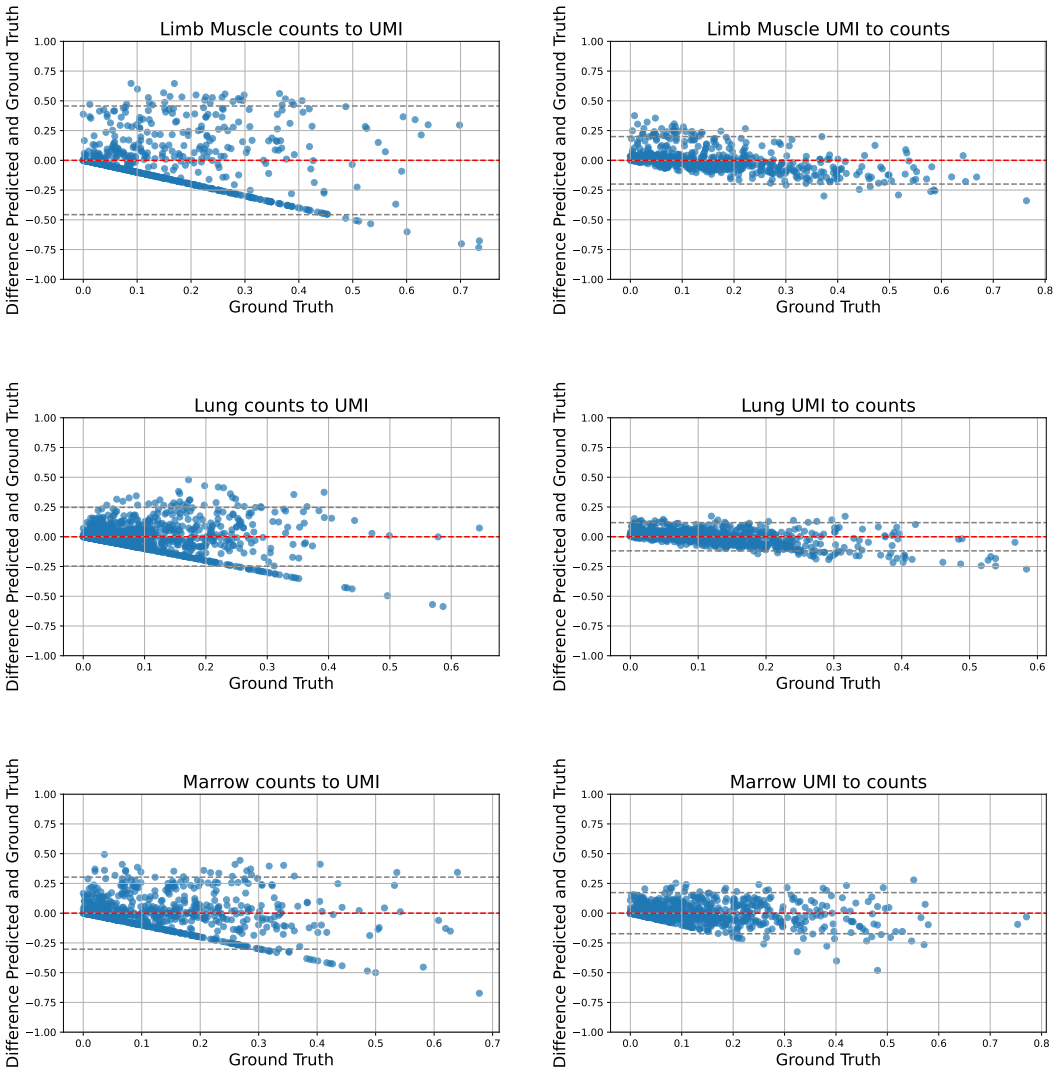
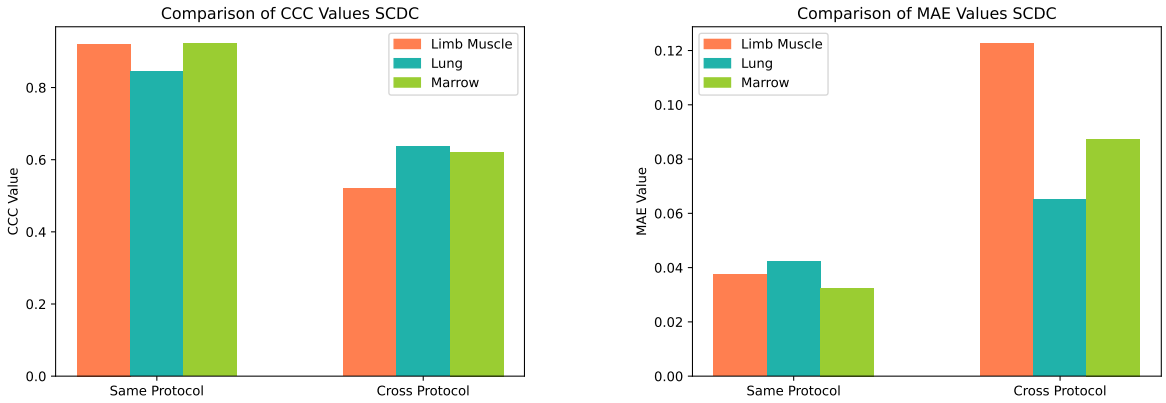


Figure 4.4: Bland-Altman plots for SCDC approach and cross-protocol scenario

based. However, in the case of the study, as we are working with proportions, homoscedasticity could not be the theoretically ideal situation, as if the variance is constant for all the proportions, the same value would not have the same impact on a small proportion as on a bigger one. Before evaluating the figures, note that y-axis limits are between -0.5 and 0.5 on the same protocol plots, while between -1 and 1 on cross-protocol plots. This is important to notice as the shape of the data is strongly influenced by the limits of the axes chosen.

Having all in mind, Figure 4.3 shows the experiments plots regarding the same-protocol scenario. If we focus on the first subplot, the first pattern to notice is an oblique line limiting the points from below. This pattern is repeated for all the subplots. This may be because we are



(a) (b)

Figure 4.5: Results on SCDC method, for CCC (a) and MAE (b) values.

working with proportions, which cannot be lower than zero value when estimating. Therefore, it will be ignored from now on when analyzing error variance, as may lead to misleadings.

Continuing with the first subplot, we can visualize a horn shape where the values are heteroscedastic, having greater differences on bigger proportions. This pattern is repeated on the fifth ('Marrow counts') subplot. Compared to the absolute error subplots of 'Limb muscle counts' and 'Marrow counts' in Figure 4.1, they show more or less constant low values of errors (corresponding to dark blue areas) which represent the beginning of the horn, while there are small high error peaks (yellow peaks), that correspond to the greater differences on bigger proportions of the end of the horn.

On the contrary, on subplots fourth ('Lung umi'), fith ('Lung umi'), and sixth ('Marrow umi'), again contrasting info of both Figures 4.3 and 4.1, the Altman plots show a homoscedastic pattern, while the absolute error plots represent constant values, smaller in the case of the last subplot. Again, both figures are consistent with each other.

Regarding the experiments on the cross-protocol scenario, represented in Figure 4.4, the first columns subplots follow a pattern of homoscedasticity again, which is consistent with the absolute error plots on Figure 4.1, representing constant values. However, it is worth noting that second column plots, corresponding to 'UMI to counts' experiment, show an oblique linear pattern that represents a bias on positive error for low proportions and negative one for larger proportions. This is shown on the absolute error as a plot with very uneven values, forming sharp peaks.

To finish with the SCDC evaluation, the CCC and MAE values were calculated for the

predicted **P** and it is ground truth for the three different tissues, and both protocols are shown in Figure 4.5. CCC comparison shows clear higher values for the same-protocol scenario, which represents a higher concordance between predicted and ground truth proportions. Furthermore, MAE values, consistently with the absolute error plots, are much lower for the same protocol scenario.

In summary, this first evaluation of the SCDC method gives the clear deduction that the method works much better for the same-protocol scenario, having much lower values of absolute error and MAE, and higher CCC. Additionally, the Bland-Altman plots show a tendency to homoscedasticity on the same-protocol scenario, having also some bias in the cross-protocol one.

4.1.2 TAPE pseudo bulk experiment

For the deep learning-based method of choice, there was no necessary processing of the data as it was already on read raw counts, so the datatype parameter was set to 'counts'. After all of the parameters were set as explained in Section 3.2.2, the experiment finishes with the deconvolution function, which outputs the needed files, saved as CSV files for later evaluation. As in the previous experiment, these steps are repeated on the three types of tissues, and changing the inputs of the function to conduct the same-protocol and cross-protocol experiments.

As before, for the evaluation of the experiment, we focus on evaluating the accuracy and reliability of predicted data of proportions of cell types (matrix **P**) by comparing it with the available ground truth of label matrix. The 3D representation of the absolute error for each scenario on TAPE can be visualized on Figures 4.6 and 4.7. Each figure represents six subplots for each of the tissues and protocols tested in the experiments. In the same-protocol scenario, the errors lie between 0 and 0.5, having some higher peaks of 0.7 error. Something similar is shown in the cross-protocol scenario, where the errors are more or less consistent with the before experiment. Although TAPE shows larger errors for the same-protocol scenario than SCDC, it is more consistent with the cross-protocol scenario.

Next, the Bland-Altman plots reveal the level of agreement between predicted and ground truth proportions. Similar to the absolute error plots, we can observe certain consistency between the Figures 4.8 and 4.9. All of the plots follow an oblique linear tendency already visualize in the cross-protocol experiment on SCDC. As explained, this tendency shows a bias toward positive error for small proportions and toward negative error for big proportions. Contrary to SCDC, which showed different patterns depending on the experiment, TAPE follows

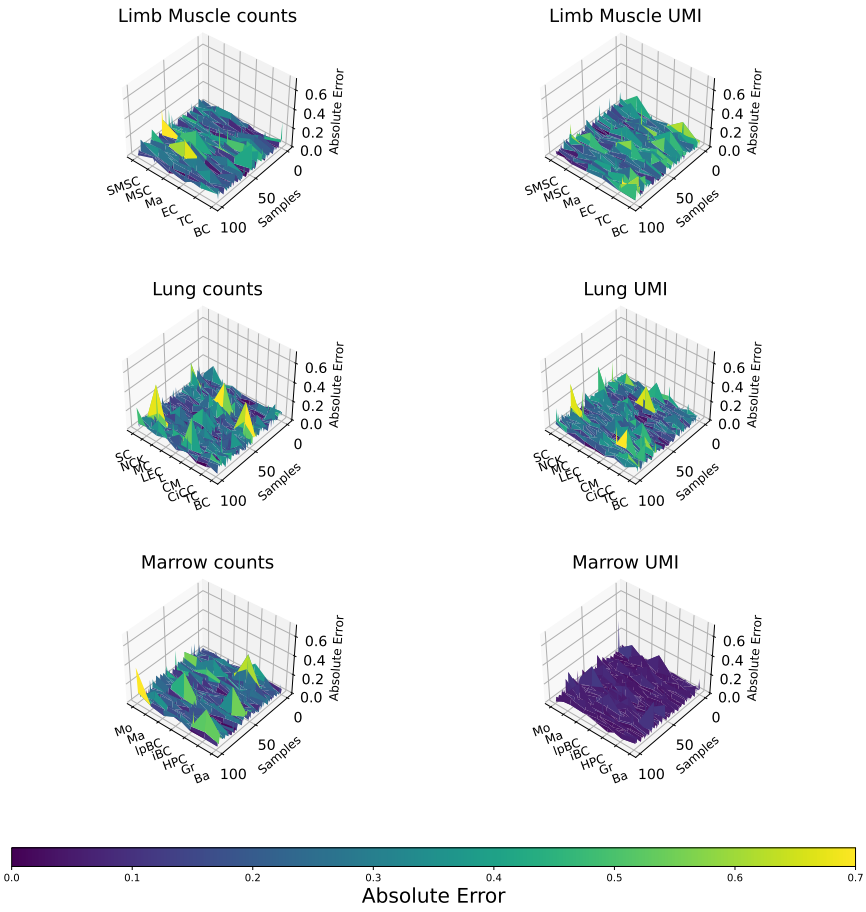


Figure 4.6: Absolute error for TAPE **P** matrix and same protocol scenario.

this pattern for all the experiments. As TAPE is a deep learning model, it is more consistent and robust across experiments, however, in this case, the results are not adequate as we are looking for heteroscedasticity. As occurs on SCDC, these figures are also consistent with the absolute error plots, Figures 4.6 and 4.7, which in general represent the sharp peaks already described.

Finally, CCC and MAE values were also calculated for TAPE experiments, resulting in the plots in Figure 4.10. Although these values also show a tendency for better performance on the same-protocol scenario, similar to SCDC, in this case, the differences are not so marked. In the case of CCC values, for instance, the value for Lung tissue on the same protocol is almost the same as the value of Limb Muscle tissue for the cross-protocol scenario. Similarly,

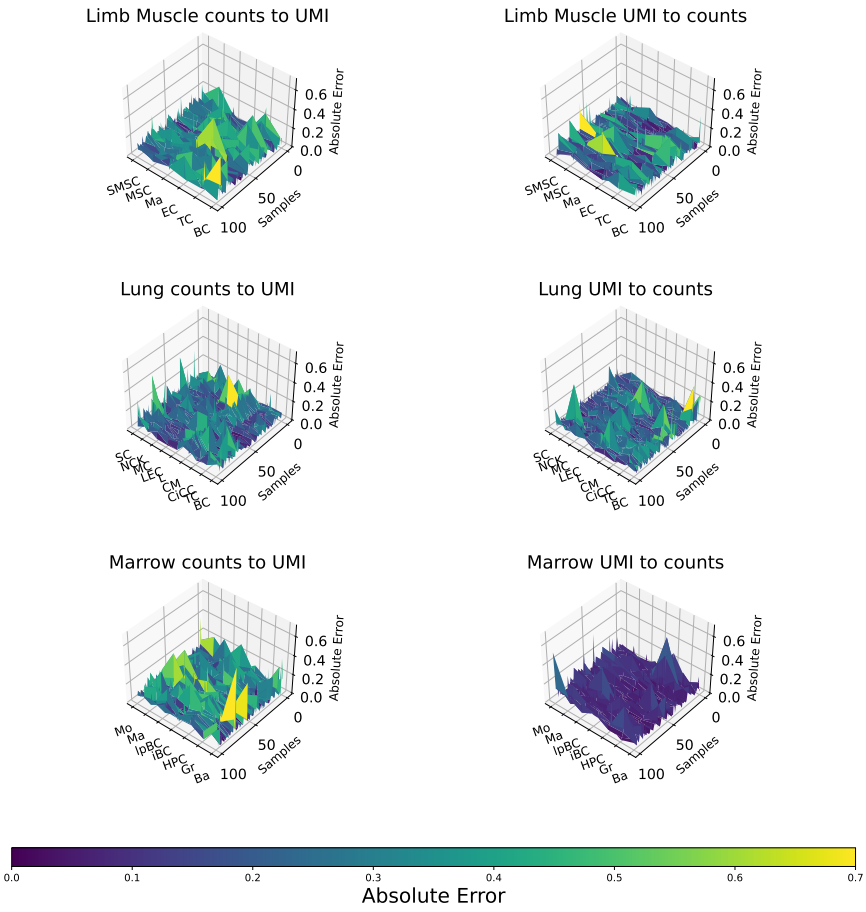


Figure 4.7: Absolute error for TAPE **P** matrix and cross-protocol scenario

all of the MAE of the same protocol experiment is near to the value of MAE for the Lung tissue experiment on cross-protocol. In general, TAPE did not work well on the Lung tissue experiment, which may be due to the fact that this is the tissue with the higher number of cells among the ones studied.

To sum up, TAPE has shown to be a more consistent method across protocols, however, the absolute error values are higher than in SCDC. Additionally, the Bland-Altman plots confirm its robustness but at the cost of showing the bias of the model towards proportion estimation, which stated this method as a non-valid deconvolution method.

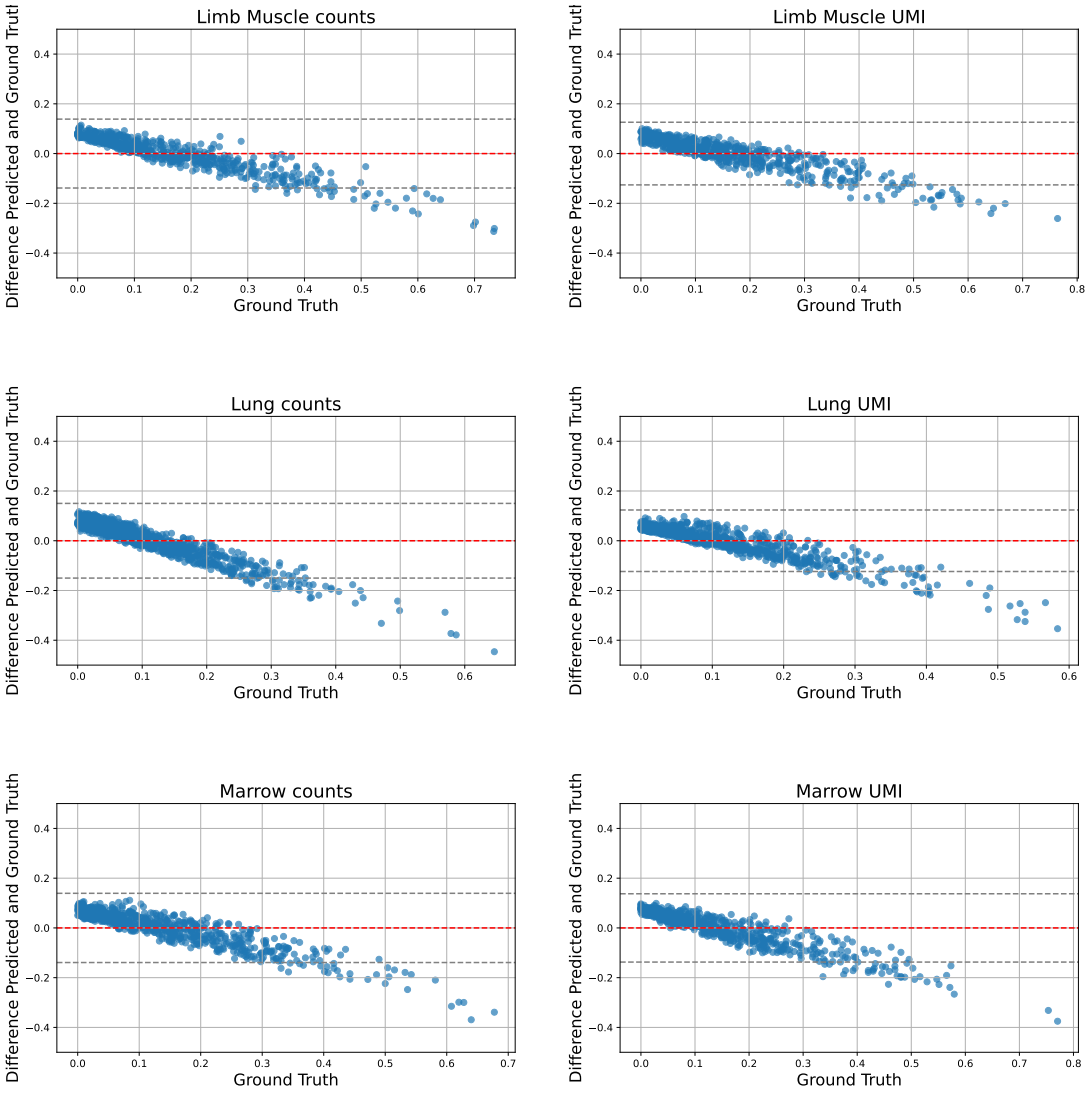


Figure 4.8: Bland-Altman plots for TAPE approach and same-protocol scenario.

4.1.3 Comparison of methodologies: TAPE vs SCDC

To compare both algorithms, we would evaluate their performance by the CCC and MAE calculations on the two approaches having in mind the protocols crossing, mixing the results of all three tissues. Then, we would focus on their best performance protocol approach to evaluate their differences across tissues. Furthermore, the differences across Bland-Altman plots would be also commented on and justified. To finish with the comparison, the results of the Wilcoxon test would be discussed.

Consequently, Figures 4.11 and 4.12 represent the CCC and MAE values as box plots where

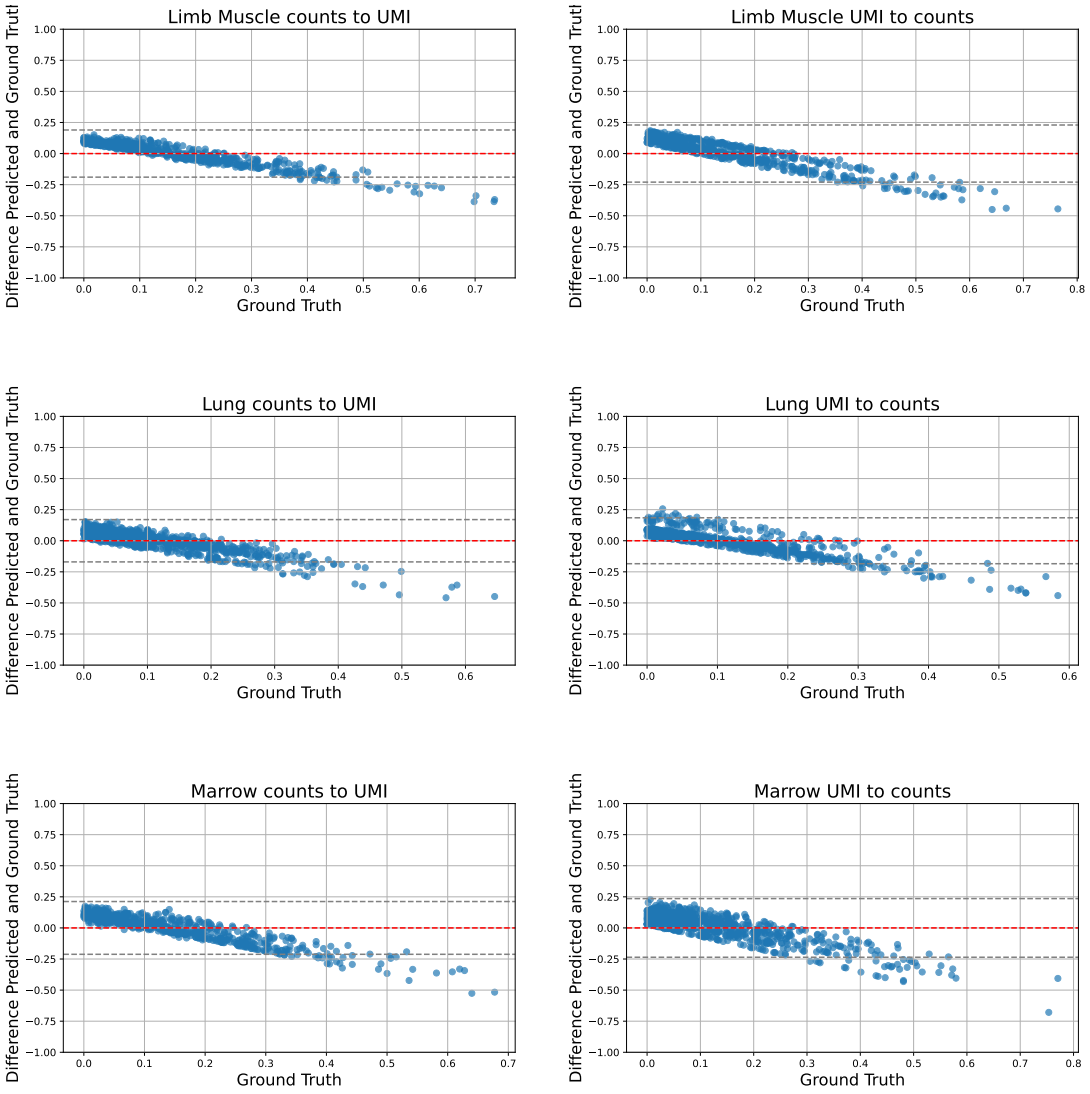
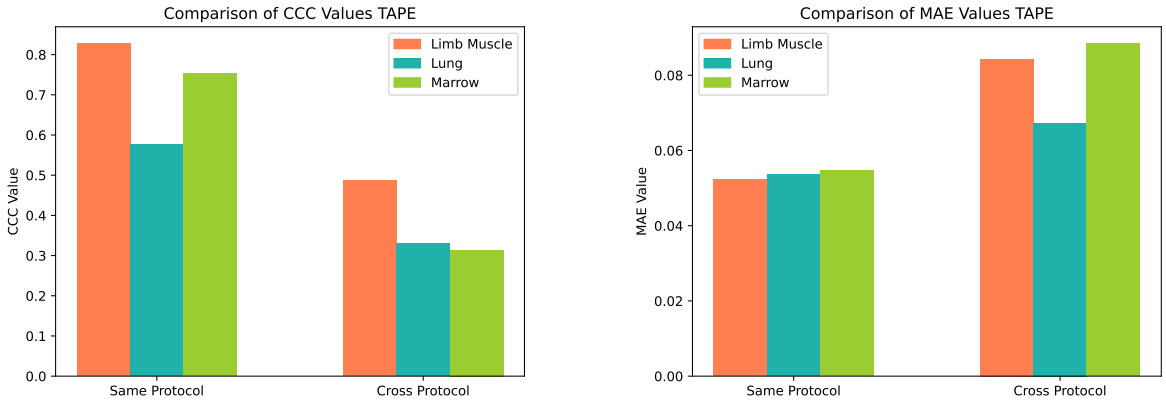


Figure 4.9: Bland-Altman plots for TAPE approach and cross-protocol scenario.

each box contains metric values for all the cell types considered in all the tissues, and the different colours refer to the different algorithms. Important to highlight the y-axis limits, that in case of CCC values vary from 0 to 1, which refers to some kind of agreement (remember that 1 indicates perfect agreement, 0 indicates no agreement, and -1 indicates perfect disagreement). In case of MAE, y-axis limits vary from 0 to 0.2, which in general are low values that indicate more or less accurate predictions.

Therefore, without considering differences across tissues experiments, the first observation is that CCC values of the SCDC method are slightly higher than on TAPE. However, it is worth noting that TAPE presents outliers on both protocol scenarios, positively, that is, higher, in the



(a) (b)

Figure 4.10: Results on TAPE method, for CCC (a) and MAE (b) values.

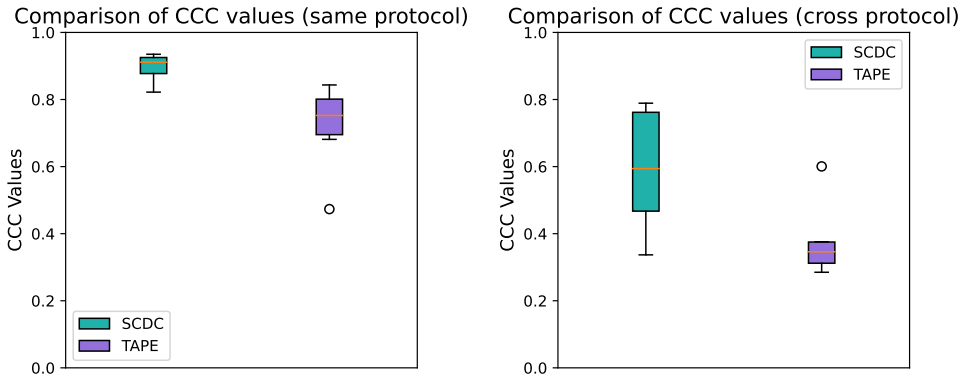


Figure 4.11: CCC values comparison between SCDC and TAPE

case of the cross-protocol scenario, and negatively in the case of the same-protocol scenario. Another observation is that TAPE maintains the variance between values on both protocols, having the same size boxes, while SCDC is much more consistent on the same-protocol than in the cross-protocol scenarios, having great variance on the last one.

Continuing with the MAE boxplots, the first observation is that, although SCDC obtains less MAE altogether, the differences between method in this case are not that as remarkable as in the case of CCC. Nevertheless, we observe again the consistency in TAPE values and the inconsistency between protocols on SCDC, having much greater variance on the cross-protocol scenario, even showing a big outlier.

As a summary, observing the general values of CCC and MAE on both methods, we can

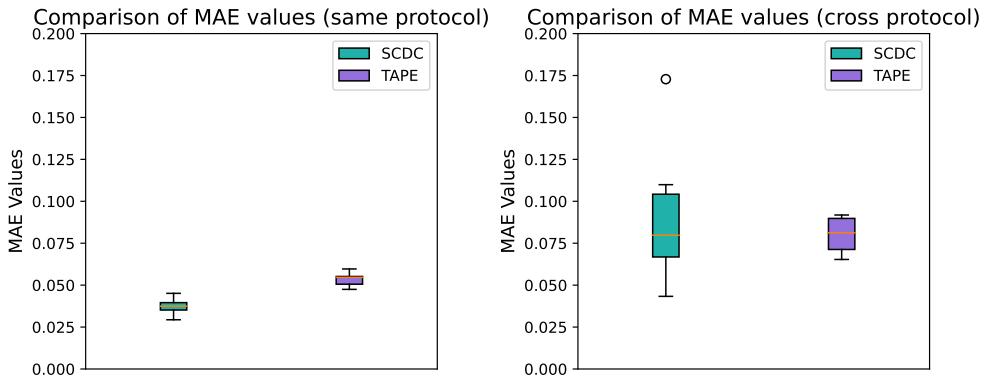


Figure 4.12: MAE values comparison between SCDC and TAPE

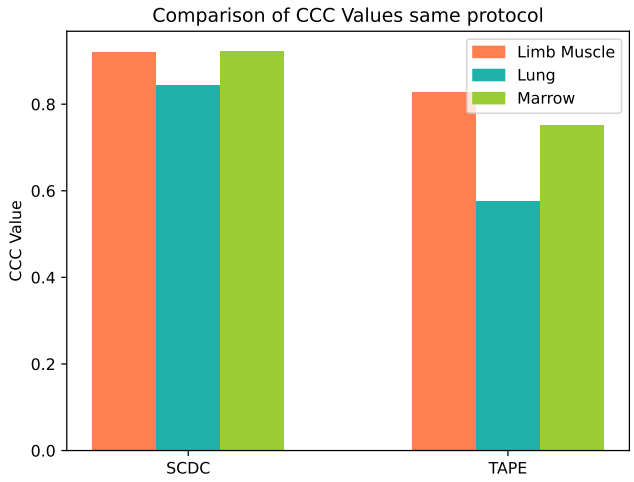


Figure 4.13: CCC values for same protocol scenario on TAPE and SCDC

conclude that, although the SCDC method works better in general, the TAPE method has the advantage of greater consistency across protocol experiments.

Leaving aside the cross-protocol scenario, where both methods have proven to have much worse performance, we would focus on the same-protocol scenario to see differences across tissues and methods. Thereafter, Figure 4.13 represents the same values of CCC but only for the same protocol approach, as has been observed to be the best performance. Can be observed in the differences across tissues, where a clear downgrade in performance is present in Lung tissue for both methods. Remembering the data described in Section 3.1.1, Lung tissue has 9 cell types, which explains the worst performance on both methods when having a greater

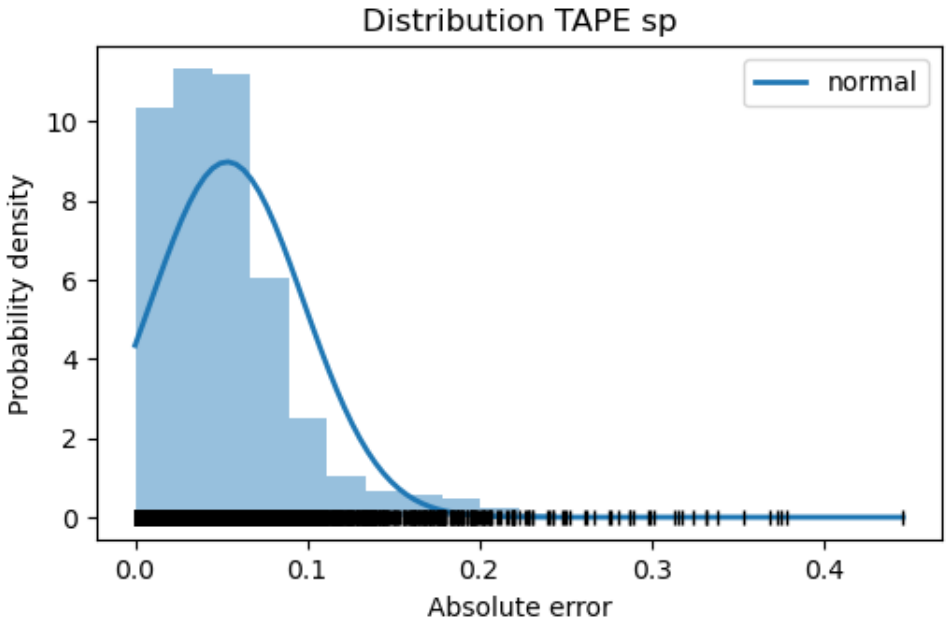


Figure 4.14: Histogram of same protocol TAPE method absolute error values.

number of cells. Specifically, in the case of TAPE, this difference across tissues is more evident, as CCC values vary clearly with the number of cell types: 0.83 for 6 cell types, 0.58 for 9 cell types and 0.75 for 7 cell types. In the case of SCDC the difference is only notable on Lung tissue. In conclusion, both methods present clear dependence on performance with the number of cell types, however, this dependence is greater on TAPE than in SCDC.

Next, differences across Bland-Altman plots between methods again reveal their great disparity. Focusing on the same-protocol scenario, Figure 4.3 of SCDC represents many differences across the subplots. Some have a tendency to heteroscedasticity while others to homoscedasticity, but in general they do not show any unusual patterns. However, Figure 4.8 of TAPE method represents a linear tendency for all the experiments that reveals a clear bias of the method.

Finally, some statistical approach was taken to evaluate the predicted proportions. Referring to the Wilcoxon test described in Section 3.2.5, this approach was taken due to the not normal distribution of the data evaluated. As an example, one of the sets of data used during this test is represented in Figure 4.14, where we can observe how the values do not follow the usual Gaussian distribution. Furthermore, the Shapiro-Wilk normality test [32] was also performed for security, resulting in a not normal distribution. Then, the p values obtained for each set of data are:

- SCDC vs TAPE same-protocol scenario: $p = 9.77 \times 10^{-100}$
- SCDC vs TAPE cross-protocol scenario: $p = 0.72$

Evaluating the given results, we can conclude that for the same protocol scenario, p value is way smaller than the established $\alpha = 0.05$, therefore, there is sufficient evidence to reject the null hypothesis and accept the alternative one, having the conclusion TAPE and SCDC absolute errors are not equal on same protocol scenario. However, in the other case the p value is greater than α, so the deduction is that in the cross-protocol scenario, both methods can be said to have similar absolute error values. Relating this to the before conclusion, both methods work poorly for the same protocol scenario.

Having all in mind, although TAPE shows robustness across protocol-changing scenarios, it is not accurate enough for deconvolution and, additionally, it shows bias on proportion estimation. On the other hand, SCDC is less consistent across protocols but has a good performance on the same-protocol scenario, which leaves SCDC as the more adequate method in terms of this experiment on pseudo bulk data.

4.2 Experiment with real data

In contrast to the experiment with pseudo bulk data, this experiment is not oriented towards ground truth comparison, as it is not available. However, the benefit of real data is that it represents real tissue expression. Then, it was decided to reproduce prior findings on the negative connection between haemoglobin A1c (HbA1c) levels, an essential biomarker for type 2 diabetes, and β cell function in order to test the effectiveness of the study techniques on real data [33] [34]. A linear model was built using the estimated cell-type proportions as the response variable and the other covariates (age, gender, BMI and HbA1c) as predictors.

The objective of this experiment is to extrapolate some of the conclusions obtained from the pseudo bulk data experiment, as the best performance of SCDC or the differences in deconvolution across the number of cells chosen. Regarding the data used, described in Section 3.1.2, the first dataset from Baron [12] consists of a total of 14 cell types, being: acinar, beta, delta, activated stellate, ductal, alpha, epsilon, gamma, endothelial, quiescent stellate, macrophage, Schwann, mast, and T cell. Secondly, the dataset from Segerstolpe [23] consists of another 14 cell types: delta, alpha, gamma, ductal, acinar, beta, unclassified endocrine, co-expression, MHC class II, PSC, endothelial, epsilon, mast, and unclassified.

Having this in mind, the common cells are chosen: alpha, beta, delta, gamma, acinar, ductal, mast, epsilon and endothelial. Mast cells were excluded due to low quantity among all. Finally, for the six cell types experiment the first six were chosen due to the high percentage of appearance, leaving a final dataset with a total of 995 cells for the Segerstolpe data, and 6517 for the Baron data. Second, for the eighth cell types experiment, all of the common cells (but the mast) were used, leaving a dataset with 1013 cells of Segerstolpe, and 6779 cells of Baron data.

To perform the experiment on SCDC, the process was simple to follow as data was already on *ExpressionSet* object type. Cell types were obtained from the experiments performed on [17], and the final files were saved as CSV. For the real data experiment on TAPE, the data was extracted from the *ExpressionSet* object to be used in Python. Then, all parameters were set as explained in Section 3.2.2 and the final files are saved as CSV files. Finally, with all the proportion matrix saved files, the regression models were constructed and plotted.

The hypotheses established for regression models obtained for the evaluation of the experiments are the following:

- Null hypothesis (H0): there is no association between β cells and HbA1c.
- Alternative hypothesis (H1): there is association between β cells and HbA1c.

If the alternative hypothesis can be evidenced by the p values obtained, that experiment would result successfully in deconvolution performance, as the negative connection between HbA1c and β cell function would have been obtained from predicted proportion values.

4.2.1 SCDC pancreatic islets experiment

To evaluate the performance of the methods further than using two different data sets, we perform another distinction between the number of cell types deconvoluted. In previous experiments on pseudo bulk data, some deductions about the worst performance when increasing the number of cell types were made, so, we would like to compare it with the real data deconvolution.

Therefore, starting with the SCDC algorithm, we observe the regression results from both datasets for the six cell types experiment in Figure 4.15. The first thing to notice in this Figure is that the negative correlation between β cell proportions and HbA1c only has been proven with the Segerstolpe dataset, obtaining a sufficient low p value. However, in the case of the Baron dataset, it failed to detect the association. If we focus on Figure 4.16 where the eighth

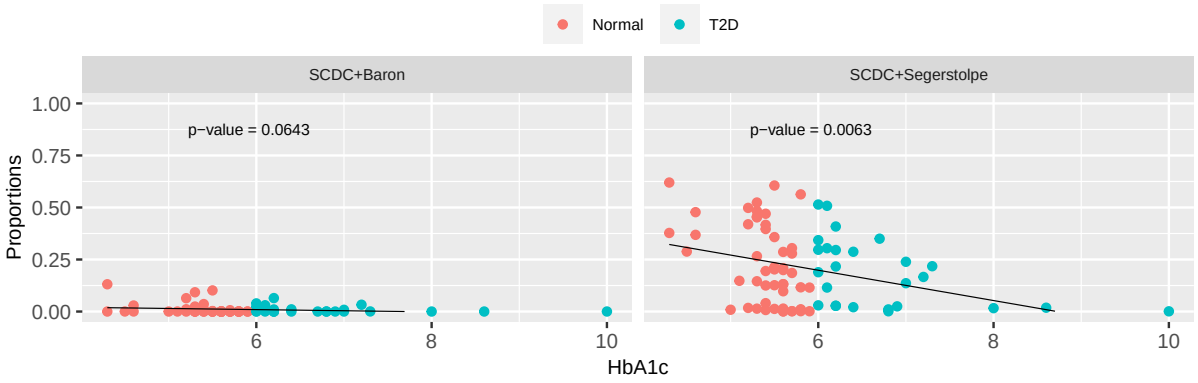


Figure 4.15: Beta cell proportions and HbA1c levels linear models for six cell types deconvolution on SCDC of two sets of real data.

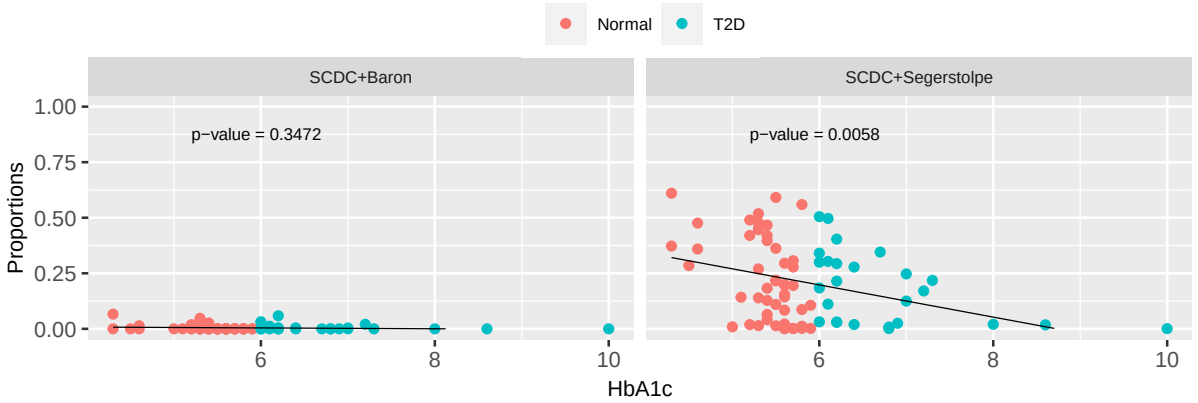


Figure 4.16: Beta cell proportions and HbA1c levels linear models for eighth cell types deconvolution on SCDC of two sets of real data.

cell types experiment is shown, this tendency is repeated. The sizes of the datasets can explain this, because, as stated at the beginning of this section, Baron dataset contains way more cell info than Segerstolpe. In the case of SCDC, the massive amount of data works against it, complicating the computation and therefore the final result of the deconvolution.

Next, focusing on the differences between the cell input size, for the Baron dataset the influence is higher than in Segerstolpe, having a way smaller p value for six cells than for the eighth. On the contrary, Segerstolpe results are similar on both cases. Therefore, SCDC is more robust in cell input size when the scRNA-seq data input size tends to be small rather than large.

In summary, the SCDC method generally has better performance and robustness when the real data inputs are not massive but medium. This can be explained by the algorithm that relies

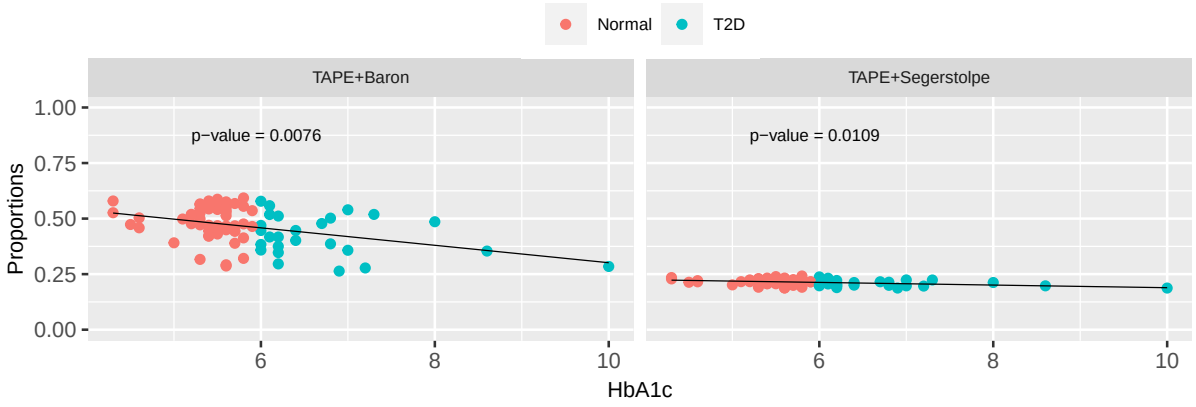


Figure 4.17: Beta cell proportions and HbA1c levels linear models for six cell types deconvolution on TAPE of two sets of real data.

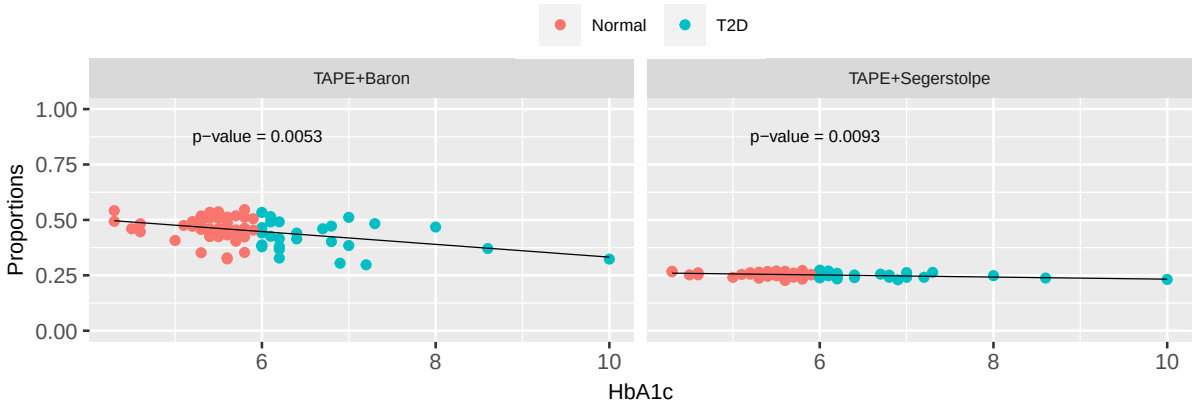


Figure 4.18: Beta cell proportions and HbA1c levels linear models for eighth cell types deconvolution on TAPE of two sets of real data.

on, W-NNLS, a statistical-based approach that works based on MVW and may get complex with large datasets.

4.2.2 TAPE pancreatic islets experiment

Continuing with the deep learning-based method, TAPE, the results of the six and eight-cell types experiment are shown in Figures 4.17 and 4.18. Contrary to the previous section, for TAPE the results on the Baron dataset are more promising. On all of these plots the association between β cell proportions and HbA1c, with all p values smaller than 0.05. Moreover, the negative correlation is more evident for the Baron dataset in both cases. Then, referring to the

dataset sizes, we can conclude that TAPE works better on massive data inputs. Furthermore, comparing the results between six and eight cells, p values are smaller in the case of eight cells, so, again, the larger the input size and the greater the number of cells, the more robust the model is and the better it performs. TAPE is a deep learning method that works by a deep network that is trained with the inputs to output a result, therefore, it is meaningful that the more data you have, the better the model trains.

4.2.3 Comparison of methodologies: TAPE vs SCDC

After the individual evaluation of both methods, we must compare their performance with each other. Focusing again on the results of the linear regression between β cell function and HbA1c, the first thing to notice is that TAPE has been able to detect the association in the four cases of the experiment, while SCDC only has been able to do it for the Segerstolpe dataset. Besides that, the mean p value for TAPE is 0.0083, while the mean p value for SCDC is 0.1059, mainly due to the bad performance on the eight cell experiment for Baron dataset, where the p value has increased until 0.3472. A plus for SCDC is that in the Segerstolpe experiments, the p values are slightly smaller than in TAPE, however, this does not compensate for its poor performance elsewhere.

In conclusion, contrary to the pseudo bulk experiments, TAPE has appear to have better performance on real data experiments. Moreover, the robustness proven with the cross-protocol experiments has been confirmed for the different datasets shown in the real data experiment, proving that TAPE is way more robust than SCDC. Regarding the cell types input size, although for SCDC the pseudo bulk deductions persist, in this case of real pancreatic islets for TAPE, the method has proven to perform better with more cells, maybe due to its deep network algorithm. Ultimately, the objective of proving the pseudo bulk experiment conclusions has met with either failures and successes.

Chapter 5

Conclusions and future lines of action

This last chapter presents the conclusions that have been reached throughout this EDP and its possible and future lines of study of deconvolution algorithms for bulk genetic expression.

5.1 Conclusions

The main objective of this EDP was to compare the two more up-to-date algorithms for bRNA-seq deconvolution, namely, SCDC and TAPE. In order to do it accurately, several types of data, scenarios and evaluation methods were needed, and they all resulted in the following conclusions.

Regarding the pseudo bulk experiments, both methods have resulted to work properly only in the same protocol scenario, leaving bad metrics results for the cross-protocol experiments. This is not a condemnation of their usage, nevertheless, many of the data utilized in this sector come from diverse protocols and can provide inefficient findings if we are not careful. Then, comparing both methods only for their best performance scenario, metrics shown as best method SCDC, having greater CCC and smaller MAE errors. In addition, SCDC showed a tendency to homoscedasticity on the difference between predicted proportions and ground truth. This, contrary to popular belief, would not be the most suited in this scenario where we are working with proportions because the differences in the large proportions should be higher than the smaller ones. We can argue that TAPE is more resilient than SCDC, yet, the differences between predicted proportions and ground truth suggest a significant bias in the model. As a result, we may conclude that SCDC performed better with pseudobulk data.

However, this conclusion is debatable due to the results obtained from real data experiments. In this occasion, SCDC failed to detect the association between β cells and HbA1c for some of the experiments, concluding that this method usually has better performance and robustness when the scRNA-seq data inputs are not massive but medium. However, the case of TAPE was that it detected the association for all types of inputs, having better performance with massive data sizes and a greater number of cell types. This confirms the robustness of the method, but it contradicts the fact that for pseudo bulk data the more cells there were, the worse the predictions were. In summary, experiments on real data have confirmed some of the conclusions of the previous ones (robustness of TAPE), but differ in others (SCDC performs worse than TAPE). Because the evaluation of these experiments on real data is not as accurate as those of pseudo bulk, not having ground truth data, more regression models or other evaluation methods would be needed to be able to faithfully confirm that TAPE is better than SCDC. Therefore, these findings do not mean that SCDC is not a suitable method for deconvolution, but that it may work better with smaller databases.

Related to the biomedical engineering competencies applied to the EDP, we can emphasize some of them, such as the ability for analytical and critical thinking, data collection and processing, evaluation of results, biotechnology and genomics knowledge, python programming, statistics evaluations etc. Furthermore, the competencies acquired during the EDP are the concept of deconvolution of bulk data, the processing of new data formats, R programming, Bland-Altman plots, the Wilcoxon test etc.

All in all, coming back to the main goal of this EDP, it is difficult to determine one of the two methods as the best one, as both of them have their advantages and disadvantages. It is clear that both need more expertise to work across protocol data, but, in general, their performance is adequate in the same protocol scenario. Numerous additional sorts of real data are required to assess their effectiveness on real-world data since the first comparative experiments between these up-to-date methods have drawn conclusions that differ from the pseudo bulk experiments. Due to the intricacy of the algorithms, a more thorough examination that incorporates every parameter that influences them might provide us with more precise findings as to which technique is superior. Nevertheless, the primary objective towards development and innovation is to take the benefits of each and apply them to the enhancement of deconvolution algorithms and the progress of genomics.

5.2 Future lines of action

The research carried out in this EDP has been restricted by various factors, including the limited computational capacity and the massive size of the data. Therefore, there are several open lines of improvement and new unexplored possibilities around this topic that are mentioned below:

- Continue with the comparison research of this EDP, having in mind more data types and protocols.
- Evaluate the compared methods on different scenarios, apart from cross-protocol, such as batch correction or noise filtering.
- Improve the algorithms for the cross-protocol scenario.
- Develop a new improved method taking into account all the advantages of both methods.

Appendix A

Real data representation

This short section shows Table A.1 where a representation of the **P** matrix from real data experiment on TAPE method is shown for a better understanding of the data used during this EDP.

	acinar	alpha	beta	delta	ductal	gamma
Sub1	0.014	0.0	0.413	0.114	0.401	0.058
Sub3	0.04	0.0	0.423	0.117	0.359	0.06
Sub6	0.0	0.0	0.587	0.147	0.22	0.045
Sub8	0.015	0.0	0.526	0.124	0.273	0.062
Sub11	0.099	0.0	0.264	0.106	0.48	0.051
Sub13	0.015	0.0	0.47	0.118	0.338	0.059
Sub14	0.038	0.0	0.321	0.108	0.475	0.058
Sub15	0.004	0.0	0.541	0.142	0.262	0.051
Sub16	0.012	0.0	0.503	0.116	0.312	0.056
Sub17	0.01	0.0	0.518	0.132	0.285	0.055
Sub18	0.013	0.0	0.472	0.115	0.343	0.057
Sub19	0.02	0.0	0.473	0.116	0.332	0.058
Sub21	0.0	0.0	0.579	0.138	0.238	0.045
Sub23	0.018	0.0	0.445	0.121	0.347	0.07
Sub24	0.014	0.0	0.459	0.114	0.361	0.054
Sub25	0.035	0.0	0.347	0.107	0.448	0.063
Sub26	0.027	0.0	0.476	0.12	0.329	0.048
Sub27	0.024	0.0	0.446	0.117	0.355	0.058
Sub28	0.007	0.0	0.536	0.133	0.279	0.046
Sub29	0.007	0.0	0.54	0.141	0.26	0.052
Sub30	0.005	0.002	0.552	0.145	0.235	0.061
Sub31	0.019	0.0	0.513	0.131	0.288	0.049
Sub32	0.048	0.005	0.354	0.121	0.408	0.065
Sub33	0.012	0.0	0.543	0.125	0.26	0.059
Sub34	0.019	0.0	0.545	0.148	0.231	0.058
Sub35	0.037	0.0	0.417	0.109	0.377	0.06
Sub36	0.041	0.0	0.391	0.112	0.397	0.059
Sub37	0.012	0.002	0.519	0.138	0.266	0.063
Sub38	0.005	0.003	0.578	0.151	0.201	0.061
Sub39	0.02	0.0	0.526	0.126	0.274	0.055
Sub40	0.11	0.0	0.287	0.108	0.443	0.053
Sub42	0.041	0.0	0.386	0.121	0.394	0.059
Sub43	0.02	0.0	0.502	0.142	0.276	0.061
Sub44	0.023	0.0	0.431	0.121	0.361	0.064
Sub45	0.009	0.0	0.562	0.162	0.21	0.057
Sub46	0.058	0.0	0.486	0.122	0.282	0.052
Sub47	0.039	0.0	0.458	0.117	0.328	0.058
Sub48	0.02	0.0	0.511	0.13	0.278	0.061
Sub50	0.008	0.0	0.538	0.132	0.272	0.049
Sub51	0.017	0.0	0.469	0.118	0.337	0.058
Sub52	0.023	0.0	0.468	0.124	0.327	0.058
Sub53	0.065	0.0	0.296	0.11	0.463	0.066
Sub54	0.034	0.0	0.447	0.113	0.345	0.061
Sub55	0.054	0.0	0.358	0.111	0.42	0.058
Sub56	0.075	0.0	0.285	0.109	0.477	0.055
Sub57	0.013	0.0	0.498	0.126	0.304	0.059
Sub58	0.009	0.0	0.499	0.12	0.319	0.053
Sub59	0.027	0.0	0.421	0.11	0.385	0.058
Sub60	0.039	0.0	0.316	0.111	0.474	0.06
Sub61	0.089	0.0	0.358	0.11	0.388	0.055
Sub62	0.003	0.004	0.555	0.152	0.227	0.058
Sub63	0.005	0.0	0.579	0.164	0.206	0.046
Sub64	0.019	0.0	0.568	0.138	0.221	0.055
Sub65	0.033	0.0	0.375	0.12	0.401	0.07
Sub66	0.013	0.0	0.477	0.116	0.344	0.051
Sub67	0.025	0.0	0.468	0.119	0.335	0.053
Sub68	0.013	0.0	0.517	0.13	0.277	0.063
Sub69	0.005	0.0	0.553	0.142	0.248	0.052
Sub70	0.018	0.0	0.449	0.125	0.346	0.062
Sub71	0.008	0.0	0.466	0.13	0.337	0.059
Sub72	0.0	0.0	0.593	0.14	0.219	0.048
Sub73	0.016	0.002	0.443	0.126	0.348	0.066
Sub74	0.049	0.0	0.292	0.113	0.482	0.065
Sub75	0.007	0.0	0.445	0.113	0.38	0.054
Sub76	0.006	0.0	0.558	0.152	0.224	0.06
Sub77	0.021	0.0	0.417	0.118	0.388	0.055
Sub78	0.006	0.0	0.519	0.132	0.28	0.064
Sub80	0.023	0.0	0.389	0.114	0.42	0.055
Sub81	0.0	0.0	0.568	0.142	0.237	0.052
Sub82	0.059	0.007	0.278	0.107	0.477	0.073
Sub83	0.037	0.0	0.402	0.111	0.399	0.052
Sub84	0.011	0.0	0.478	0.118	0.339	0.054
Sub85	0.009	0.0	0.465	0.123	0.347	0.057
Sub86	0.01	0.0	0.529	0.135	0.27	0.057
Sub87	0.031	0.003	0.384	0.123	0.396	0.063
Sub88	0.003	0.004	0.575	0.177	0.179	0.062
Sub89	0.004	0.0	0.566	0.149	0.227	0.054

Table A.1: Example **P** matrix from real data experiment on TAPE method, using Baron dataset and six cell types.

Bibliography

[1] Xiaoqing Yu et al. “Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments”. In: *Methods in Molecular Biology (Clifton, N.J.)* 2194 (Jan. 2021), pp. 143–175. ISSN: 1940-6029. DOI: 10.1007/978-1-0716-0849-4_9.

[2] Kai Kang et al. “CDSeq: A Novel Complete Deconvolution Method for Dissecting Heterogeneous Samples using Gene Expression Data”. In: *PLOS Computational Biology* 15 (Dec. 2019), e1007510. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007510.

[3] Yanshuo Chen et al. “Deep Autoencoder for Interpretable Tissue-Adaptive Deconvolution and Cell-Type-Specific Gene Analysis”. In: *Nature Communications* 13 (Nov. 2022), p. 6735. DOI: 10.1038/s41467-022-34550-9.

[4] Jared M. Churko et al. “Overview of High Throughput Sequencing Technologies to Elucidate Molecular Pathways in Cardiovascular Diseases”. In: *Circulation Research* 112 (June 2013), pp. 1613–1623. DOI: 10.1161/circresaha.113.300939.

[5] Sampled. *Bulk RNA Sequencing vs. Single Cell RNA Sequencing – what’s the difference between these powerful techniques?* https://sampled.com/bulk-rna-sequencing-vs-single-cell-rna-sequencing/. [Accessed 20-Mar-2023]. Mar. 2023.

[6] Fuchou Tang et al. “mRNA-Seq Whole-Transcriptome Analysis of a Single Cell”. In: *Nature Methods* 6 (May 2009), pp. 377–382. ISSN: 1548-7105. DOI: 10.1038/nmeth.1315.

[7] Youjin Hu et al. “Single Cell Multi-Omics Technology: Methodology and Application”. In: *Frontiers in Cell and Developmental Biology* 6 (Apr. 2018). ISSN: 2296-634X. DOI: 10.3389/fcell.2018.00028.

[8] Xiliang Wang et al. “Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2”. In: *Genomics, Proteomics & Bioinformatics* 19 (Apr. 2021), pp. 253–266. ISSN: 1672-0229. DOI: 10.1016/j.gpb.2020.02.005.

[9] Allon M Klein et al. “Droplet Barcoding for Single Cell Transcriptomics applied to Embryonic Stem Cells”. In: *Cell* 161 (May 2015), pp. 1187–1201. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.04.044.

[10] lizard.bio. *Single-cell vs. bulk sequencing: which one to use when?* https://lizard.bio/blog/single-cell-vs-bulk/. [Accessed 20-Mar-2023]. Aug. 2021.

[11] Geng Chen. “Editorial: Multimodal and Integrative Analysis of Single-Cell or Bulk Sequencing Data”. In: *Frontiers in Genetics* 12 (Feb. 2021). ISSN: 1664-8021. DOI: 10.3389/fgene.2021.658185.

[12] Maayan Baron et al. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure”. In: *Cell systems* 3 (Oct. 2016), 346–360.e4. ISSN: 2405-4712. DOI: 10.1016/j.cels.2016.08.011.

[13] Xuran Wang et al. “Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference”. In: *Nature Communications* 10 (Jan. 2019), p. 380. ISSN: 2041-1723. DOI: 10.1038/s41467-018-08023-x.

[14] Aaron M. Newman et al. “Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry”. In: *Nature Biotechnology* 37 (July 2019), pp. 773–782. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0114-2.

[15] Daphne Tsoucas et al. “Accurate Estimation of Cell-Type Composition from Gene Expression Data”. In: *Nature Communications* 10 (July 2019), p. 2975. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10802-z.

[16] Brandon Jew et al. “Accurate Estimation of Cell Composition in Bulk Expression Through Robust Integration of Single-Cell Information”. In: *Nature Communications* 11 (Apr. 2020), p. 1971. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15816-6.

[17] Meichen Dong et al. “SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References”. In: *Briefings in Bioinformatics* 22 (Jan. 2021), pp. 416–427. ISSN: 1477-4054. DOI: 10.1093/bib/bbz166.

[18] Kevin Menden et al. “Deep Learning–based Cell Composition Analysis from Tissue Expression Profiles”. In: *Science Advances* 6 (July 2020), eaba2619. DOI: 10.1126/sciadv.aba2619.

[19] Aaron M. Newman et al. “Robust Enumeration of Cell Subsets from Tissue Expression Profiles”. In: *Nature Methods* 12 (May 2015), pp. 453–457. ISSN: 1548-7105. DOI: 10.1038/nmeth.3337.

[20] Jürgen Schmidhuber. "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61 (Jan. 2015), pp. 85–117. ISSN: 08936080. DOI: 10.1016/j.neunet.2014.09.003.

[21] Tabula Muris Consortium et al. "Single-cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris". In: *Nature* 562 (Oct. 2018), pp. 367–372. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0590-4.

[22] Damian Smedley et al. "The BioMart Community Portal: an Innovative Alternative to Large, Centralized Data Repositories". In: *Nucleic Acids Research* 43 (July 2015), W589–W598. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv350.

[23] Åsa Segerstolpe et al. "Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes". In: *Cell Metabolism* 24 (Oct. 2016), pp. 593–607. ISSN: 15504131. DOI: 10.1016/j.cmet.2016.08.020.

[24] João Fadista et al. "Global Genomic and Transcriptomic Analysis of Human Pancreatic Islets reveals Novel Genes Influencing Glucose Metabolism". In: *Proceedings of the National Academy of Sciences* 111 (Sept. 2014), pp. 13924–13929. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1402665111.

[25] *SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References*. https://meichendong.github.io/SCDC/articles/SCDC.html. [Accessed 5-Jun-2023].

[26] Donghui Chen and Robert J. Plemmons. "The Birth of Numerical Analysis". In: Nov. 2009. Chap. Nonnegativity Constraints in Numerical Analysis, pp. 109–139. DOI: 10.1142/9789812836267_0008.

[27] *scTAPE*. https://sctape.readthedocs.io/. [Accessed 12-May-2023].

[28] L. I. Lin. "A Concordance Correlation Coefficient to Evaluate Reproducibility". In: *Biometrics* (Mar. 1989), pp. 255–268. ISSN: 0006-341X. DOI: https://doi.org/10.2307/2532051.

[29] Haldun Akoglu. "User's Guide to Correlation Coefficients". In: *Turkish Journal of Emergency Medicine* 18 (Aug. 2018), pp. 91–93. ISSN: 2452-2473. DOI: 10.1016/j.tjem.2018.08.001.

[30] J. Martin Bland and DouglasG Altman. "Statistical Methods for Assessing Agreement between two Methods of Clinical Measurement". In: *The Lancet* (Feb. 1986), pp. 307–310. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(86)90837-8.

[31] Frank Wilcoxon. “Individual Comparisons by Ranking Methods”. In: *Biometrics Bulletin* 1 (1945), pp. 80–83. ISSN: 0099-4987. DOI: 10.2307/3001968.

[32] S. S. SHAPIRO and M. B. WILK. “An Analysis of Variance Test for Normality (Complete Samples)”. In: *Biometrika* 52 (Dec. 1965), pp. 591–611. ISSN: 0006-3444. DOI: 10.1093/biomet/52.3-4.591.

[33] Mustafa Kanat et al. “The Relationship Between β-Cell Function and Glycated Hemoglobin”. In: *Diabetes Care* 34 (Apr. 2011), pp. 1006–1010. ISSN: 0149-5992. DOI: 10.2337/dc10-1352.

[34] Xinguo Hou et al. “Relationship of Hemoglobin A1c with β Cell Function and Insulin Resistance in Newly Diagnosed and Drug Naive Type 2 Diabetes Patients”. In: *Journal of Diabetes Research* 2016 (2016), pp. 1–6. ISSN: 2314-6745, 2314-6753. DOI: 10.1155/2016/8797316.