THE INSTITUTE OF MEDICAL SCIENCE, THE UNIVERSITY OF TOKYO

Laboratory of Functional Analysis in silico

Internship report

Inferring cell proportion in bulk dataset using single cell as reference

Supervisor: Kenta NAKAI, Sung-Joon PARK and Luis Augusto Eijy NAGAI

Master of molecular bioinformatic methods and analyses, University of Lyon 1

UNIVERSITÉ DE LYON

Lyon 1

Rémi SERAPHIN
29/05/2020

Abstract

RNA sequencing is now an established technology and has been widely used. However, assays are subject to cell composition as they are unable to capture cell heterogeneity present in samples. Computational method called deconvolution has been developed to estimate this composition. This method usually requires gene expression profile and cell type expression profile to be precise. In parallel, RNA sequencing technologies have been enhanced leading to the apparition of single cell RNA-seq. Single cell RNA-seq is able to target individual cells thus capturing cell heterogeneity, but is limited by its cost, depth and presence technical noise. Since both technologies bring different information, it has been recently thought to use single cell derived cell type expression profile to deconvolve related RNA-seq assay. Here we study elaboration of such computational method.

Table of contents

Table of figures

List of tables

List of abbreviations

RNA: Ribonucleic acid

RNA-seq: RNA sequencing

scRNA-seq: Single cell RNA sequencing

ERCC: External RNA control consortium

UMI: Unique molecular identifier

DNA: Deoxyribonucleic acid

cDNA: Complementary deoxyribonucleic acid

eQTLs: Expression quantitative trait loci

BSS: Blind source separation

LS: Least square

OLS: Ordinary east square

LLS: Linear least square

NNLS or NLS: Non-negative least squares

PCA: Principal components analysis

tSNE: T-stochastic neighbor embedding

List of software

Git 2.26.2

Python 3.8.3

- NumPy 1.18.0
- Matplotlib 3.2.1
- Pandas 1.0.3
- SciPy 1.4.1

R 3.6

- Seurat 3.0
- SC3 3.11
- Tidyverse 1.3.0

1) Introduction:

1.1) Context

A recent revolution in the field of biology is the advent of new sequencing technologies, called next-generation sequencing (NGS). NGS technologies allow unprecedented access to the genetic information contained in cells. Since their apparition, these technologies have been enhanced to expand their field of application. Indeed, firstly developed in the genomic area, application ranges now from transcriptomic to proteomic passing by epigenetic. In addition to the expansion of possible application, the required quantity of starting material has been reduced rendering the possibility to target individual cells. Methods targeting individual cells are called “single cell” by opposition to those that target populations of cells which are called “bulk”. The apparition of single cell RNA-seq (scRNA-seq) technologies allows to bring new light on many longstanding questions, but simultaneously raises several inherent technical issues that need to be considered. Since bulk and single-cell technologies have their intrinsic characteristics, they capture genomic information complementary to each other. Here, we focus on the application of both technologies in the field of transcriptomics.

1.2) Bulk RNA-seq

The standard RNA sequencing (RNA-seq) consists of extracting a population of RNA molecules, isolating a subset of specific RNAs by using an adapted protocol (e.g. poly-A tailed RNA), converting those to complementary DNA (cDNA), preparing the sequencing library, and sequencing the library and generating reads by NGS machines [1]–[3]. Each step should be set up based on the research context and goals. Using the RNA-seq reads obtained, the typical analytic workflow performs a quality assessment, alignment to a reference genome or transcriptome, *de novo* or reference-based assembly into transcripts and quantification of gene expression by counting reads aligned to transcripts or exon (Figure 1) [4]. The expression values from the bulk RNA-seq assay facilitate downstream analyses that focus on multiple topics: differential expression analysis, detection of allele-specific expression, identification of expression quantitative trait loci (eQTLs)…

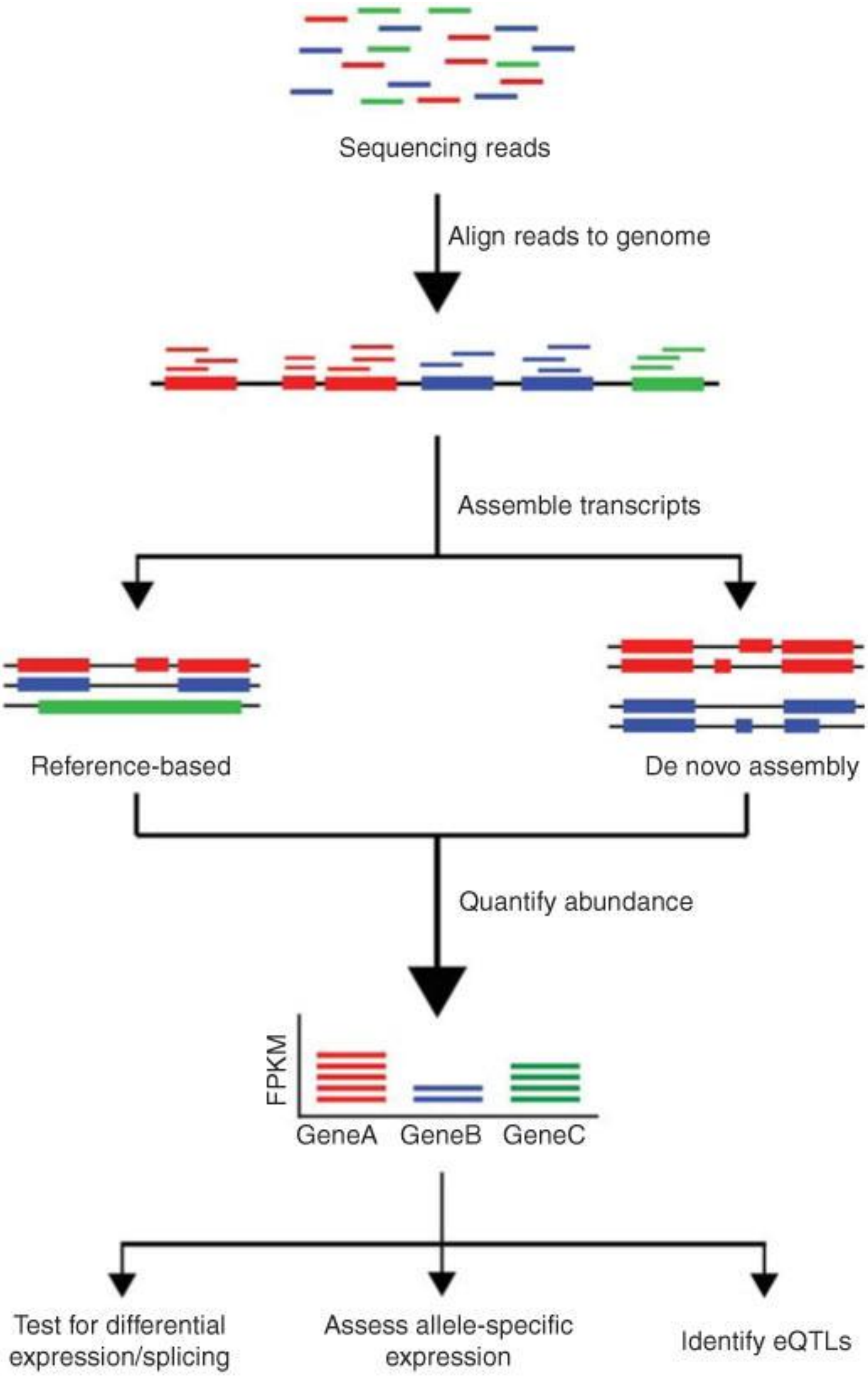


Figure 1: Overview of bulk RNA-seq analysis.

Following sequencing, reads are aligned to a reference genome, then reads are assembled into transcripts using the reference annotations or de novo assembly. Next, the expression level of each gene is estimated by counting the number of reads aligned to each feature further used in downstream analyses. (Original graphic from Kukurba et al., 2015)

A major issue on the bulk RNA-seq assay is that the assay measures averaged expression of genes in a cell population. Thus, this approach is inadequate to capture the heterogeneity often present in cell populations. Indeed, biological samples usually correspond to tissue isolated from an organism, or cellular cultures, which are by nature heterogeneous. This heterogeneity comes down to variation in expression level of each individual gene between different cells including the presence of different cell types. Thus, the cell-type diversity present in cells to be sequenced is one of the confounding factors that lead to misinterpretation. In the analysis of differentially expressed genes, highly variable genes in minor but important cell populations will be hard to detect due to the masking effect by stable expression levels in major cell populations [5]. For example, since pathological or tumoral tissue are often composed of variable proportion of malignant or disease-states cells neighboring normal cells, it is expected that those malignant cells assayed by the bulk RNA-seq will be masked. To address this issue multiple experimental methods have been developed to isolate distinct cells such as laser-capture microdissection or cell purification and enrichment. The methods have been enhanced, and their use coupled with RNA-seq have led to single-cell RNA-seq technologies.

1.3) Single cell RNA-seq

A standard single cell RNA-seq (scRNA-seq) experiment consists of isolating single cells, lysing the isolated cells in a way that conserves RNA, capturing subset of RNA molecules (e.g. poly-A-tailed RNA), converting the molecules to complementary cDNA, one or multiple steps of amplification, library preparation (usually inserting a tag to identify each origin), pooling libraries and sequencing [6], [7]. Each step should be adjusted to fit the research objectives (Figure 2). Like in the bulk RNA-seq, the following bioinformatics analyses are necessary: quality control, mapping, expression quantification. Unlike in the bulk RNA-seq, the single-cell RNA-seq analysis involves normalization, sometimes imputation, dimension reduction, and feature selection (Figure 3) [8]–[11]. The normalization goal is to extract real signal from technical noise [12]. Imputation focuses on correcting dropout events and missing values which increase variability. Both usually result from failed capture or amplification of the original RNA molecule. Dimension reduction and feature selection are used to assess the high dimensionality of single cell studies, originating from the high number of genes and cells, by focusing on key factors or meaningful genes. The scRNA-seq data are used for various downstream analyses: differential gene expression, inferring lineage differentiation, constructing gene regulatory network, detecting marker genes among other topics.

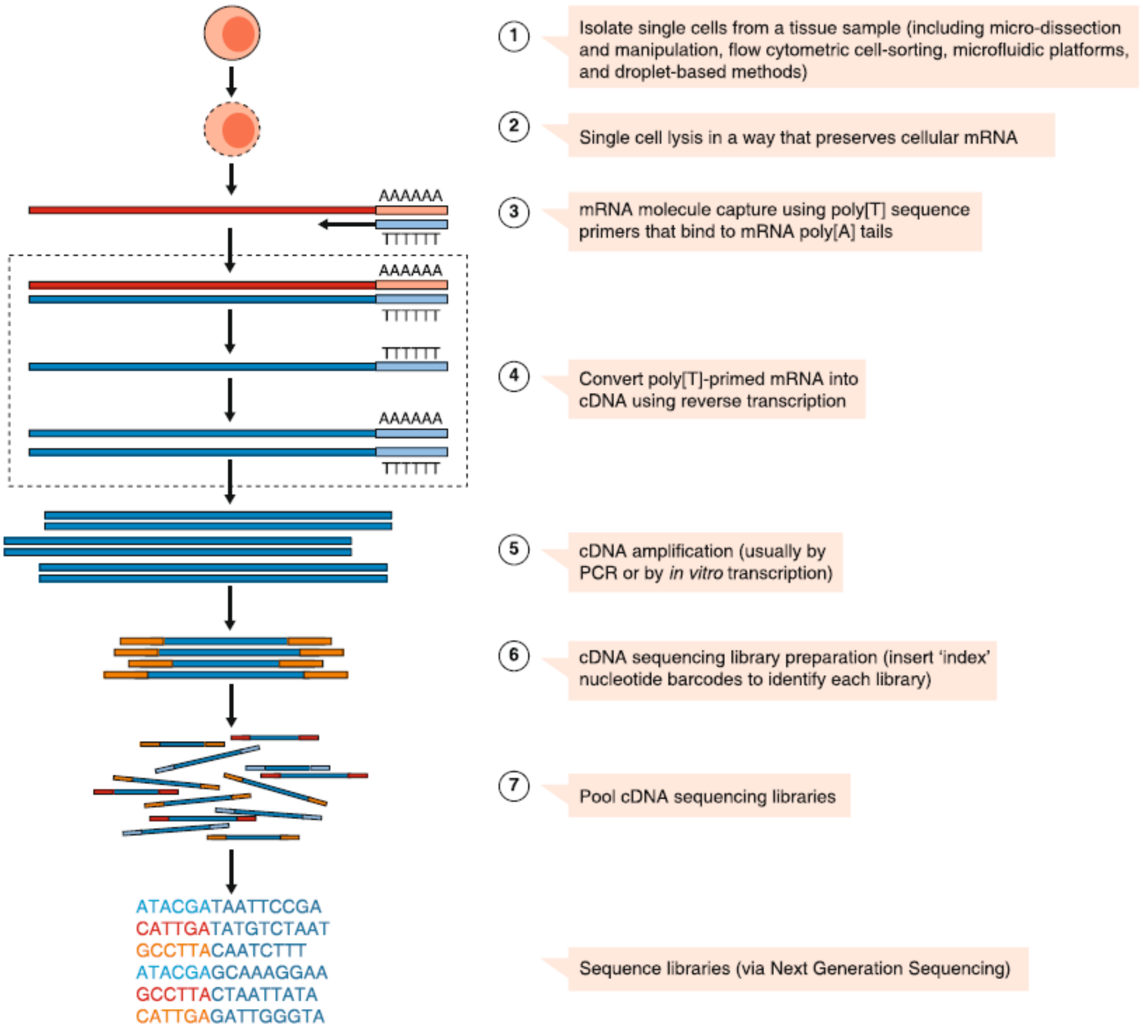

1
Isolate single cells from a tissue sample (including micro-dissection and manipulation, flow cytometric cell-sorting, microfluidic platforms, and droplet-based methods)
2
Single cell lysis in a way that preserves cellular mRNA
AAAAAA
TTTTTT
3
mRNA molecule capture using poly[T] sequence primers that bind to mRNA poly[A] tails
AAAAAA
TTTTTT
TTTTTT
AAAAAA
TTTTTT
4
Convert poly[T]-primed mRNA into cDNA using reverse transcription
5
cDNA amplification (usually by PCR or by *in vitro* transcription)
6
cDNA sequencing library preparation (insert 'index' nucleotide barcodes to identify each library)
7
Pool cDNA sequencing libraries
ATACGATAATTCCGA
CATTGATATGTCTAAT
GCCTTACAATCTTT
ATACGAGCAAAGGAA
GCCTTACTAATTATA
CATTGAGATTGGGTA
Sequence libraries (via Next Generation Sequencing)

Figure 2: Overview of a standard single cell RNAseq experiment.

1) cell isolation, 2) cell lysis, 3) mRNA capture, 4) reverse transcription into complementary DNA (cDNA), 5) cDNA amplification, 6) library preparation, 7) pooling of sequences libraries. (Adapted from Haque et al., 2017)

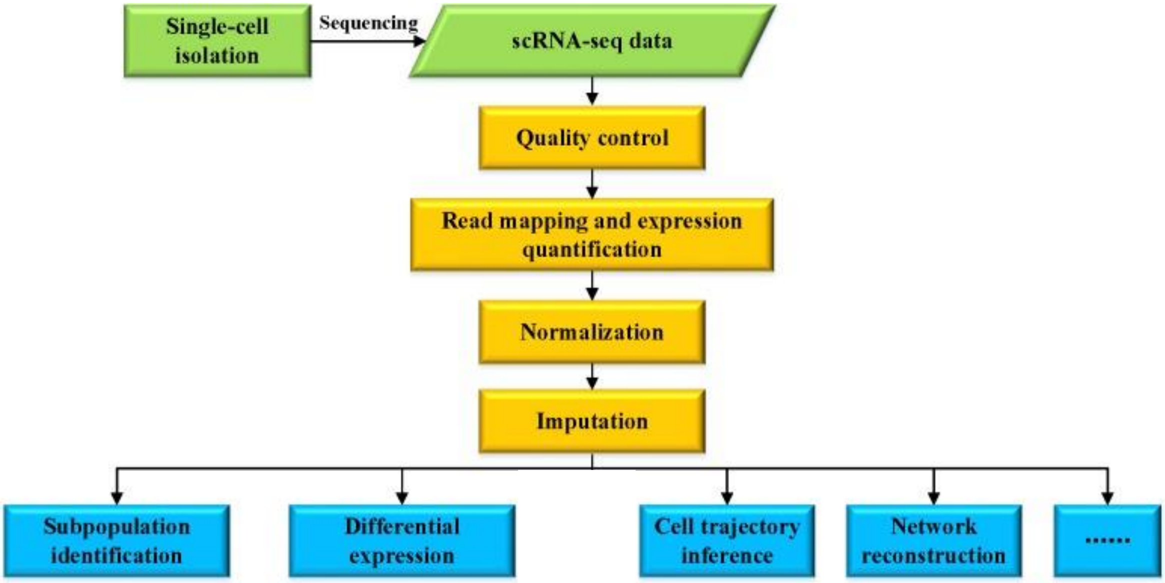


Figure 3: Overview of the workflow of a single cell RNA-seq analysis and potential downstream analyses.

(Adapted from Chen et al., 2019)

The main advantage of single-cell sequencing is that the assay allows to access the heterogeneity which bulk RNA-seq cannot achieve. Moreover, the assay is able to capture subpopulation and even cell-to-cell variability. As technical issues in the single-cell sequencing, the low amount of initial RNA molecules leads to capture failure and artifacts coming from biased amplification [13], [14]. To address those issues, protocols include in sample controls such as ERCC (External RNA Control Consortium) spike-in and Unique Molecular Identifiers (UMI). ERCC consists of adding known and controlled mRNA allow assessing the level of technical variability. UMI are randomly generated molecular tags which give indication on the original molecule of the read thus allowing reduction of amplification bias. The previously explained single cell specific methods also try to tackle technical noise but no clear consensus has been found in their uses. Also, scRNA-seq depth is mediocre in comparison to bulk RNA-seq whereby the cost of the former is elevated.

1.4) Deconvolution

To take into accounts the confounding factor that is cell type proportion in bulk RNA-seq studies, several computational approaches have been developed for estimating the proportion of cell types in a given sample [15]–[20]. This process is called computational deconvolution. The latter has application in multiple fields and is not specifically restricted to RNA-seq for the analysis of expression data from heterogeneous samples. Deconvolution can be defined as "*source separation*" and has been first used with audio signals that are a mixture of different sounds.

In biology, although it is used in image processing and fluorescence measurements, the process was first applied to array-based gene expression profiling in 2001 and later used for RNA-seq data by continuously updating the process [21]. In the case of separating signals in a dataset of gene expression from complex tissue, both cell-type specific expression and proportion are unknown. This corresponds to a situation where both the sources signal and the mixing process are unknown which make it falls under the blind source separation (BSS) category. However, when one of them is used to compute the other the case relates to a guided-BSS, also referred as partial deconvolution by opposition to full deconvolution [22]. In most cases, it is difficult to experimentally measure both mixed-expression profile and cell-type expression profile (or proportion) at the same time, rendering partial deconvolution situational.

1.5) Problematic

Recently propositions have been made to use scRNA-seq ability to identify cell populations and to define cell-type specific markers as a base to infer cell composition of related bulk sample [23]–[26]. Thus, to profit from advantages of both sequencing methods and to be able to reanalyze the wealth of publicly available bulk datasets, developing a deconvolution method using scRNA-seq derived cell-type specific markers as input to infer cell presence and proportion in a bulk dataset is of high interest.

2) Material and Methods

2.1) Deconvolution problem

In the case of gene expression in a heterogeneous sample, the expression of a gene can be modelled as the sum of expression of this gene in each cell-type weighted by the abundance of each of those cell-types. This model can be algebraically written as the following equation:

$$T_i = \sum_{k=1}^{K} C_{ik} \, . P_{k(T)} + \; e_i \; ; with \; i \; ranging \; from \; 1 \; to \; M \qquad (1)$$

Where T_i corresponds to the expression value of gene i in the sample T; C_{ik} corresponds to the expression value of gene i in cell-type k; $P_{k(T)}$ the proportion of cell-type k; e_i the error term; K the number of cell-types and M the number of genes.

Which can also be formulated as the following matrix equation:

$$T = C\,.\,P \qquad (2)$$

Where *T* corresponds to the measured expression of each gene from a complex tissue or sample; *C* corresponds to the cell-type specific expression of each gene and *P* the proportion of each cell-type in the sample *T*.

Several mathematical approaches are used to solve the deconvolution problem. The most frequently used is the group of Least Square (LS) methods also called Ordinary Least Square (OLS) or Linear Least Square (LLS), which minimizes the sum of squared residuals that represents the difference between an observed value and a fitted model. In the case of equation (2), the least square method would minimize the sum of squares of the differences between the fitted model ($\mathrm{C} \cdot \mathrm{P}$) and observed values (T) by finding the best a proportion vector (P) as follows:

$$Given\ T\ and\ C\colon min_P||C.P - \ T||^2 \qquad (3)$$

The equation (3) corresponds to an unconstrained optimization problem. However, because of the biological context of this deconvolution, two constraints can be added: 1) *non-negativity*: finding a negative number of cells would be meaningless, 2) *sum-to-one*: the sum of the proportion must be equal to one. With these constraints, the equation (3) becomes a constrained optimization problem, and the most adapted least square methods are Non-Negative Least Squares (NNLS or NLS). The application of the second constraint is applied after minimization in some cases. NNLS approaches have already been implemented in the programming language R and Python.

Other methods to solve the minimization problem include maximum likelihood estimation if the hypothesis that the error terms follow a normal distribution, simulated annealing that is a probabilistic technique, and support vector machine with the support-vector regression method that is a machine learning approach.

2.2) Programming language R

The R language serves an environment thought for statistical computing and graphics [27]. As a high-level programming language, R supports multiple programming paradigms including array, object-oriented, and functional programming. Moreover, R proposes a variety of statistical and graphical techniques extensible through user-created packages. Therefore, R is widely used in statistics and data mining fields, and it becomes popular more and more in bioinformatics as demonstrated by Bioconductor that provides tools for genomic data analysis and comprehension.

2.2.1) Seurat

Seurat is an R package that focuses on single-cell NGS data analysis, such as quality control, normalization, visualization, etc. It is thought that Seurat is standard software, as easy-to-use, clarity, and interpretability.

As a particular advantage in Seurat, it provides a versatile workflow that notably allows to cluster cells and to identify cell markers. The first step is to generate a Seurat object from a raw single-cell RNA-seq count matrix. Then, the object is filtered by a quality control process for removing unwanted cells, such as doublet or dead cells. Next, the filtered data is normalized and used for finding and selecting features (highly variable features are considered biologically interesting). Afterward, scaling is applied so that over-expressed genes do not dominate downstream analyses; SCTransform in Seurat performs the normalizing, scaling, and selecting features. Next, the dimension reduction is performed with the PCA method. Note that it is important to determine the dimensionality of the dataset to keep most of the relevant signals. Using the information of PCA, single cells are clustered. Following the clustering, differentially expressed features of each cluster are then extracted as cluster markers. Each step needs to be adjusted to the dataset. This workflow has been used to identify cell populations and cell-type-specific markers for the deconvolution. As a cluster corresponds to a cell-type in general, it is thought that the average expression of cluster markers is representative of a cell-type-specific expression.

2.3) Python

Python is also a high-level object-oriented computer programming language that provides high readability of its codes. Python has various standard libraries that make it popular in a wide range of domains including data science and bioinformatics. SciPy is a typical library of Python, which enhances its usability.

2.3.1) SciPy

SciPy is a Python library used for scientific computing [28]. It builds on the NumPy package which was built to facilitate the use of large and multi-dimensional arrays and matrices in Python. SciPy includes multiple modules for common tasks in science and engineering. The SciPy optimize module is very useful for deconvolution. Indeed, it provides function to minimize objectives function as well as least-squares solver. Thus, the nnls function allows to solve non-negative least square problem as it is a wrapper to a Fortran (programming language) solver. However, it doesn't support the sum-to-one constraint, nonetheless it can be useful as it can be used to get an initial guess that other methods require. This is the case for the more general minimize function which can also be use in the deconvolution context. Indeed, it allows to minimize the objective function given in parameter. It also allows to implement bounds and constrained through the use of different solver also called method. Both the SLSQP (Sequential Least SQuare Programming) and trust-constr allow the constrained minimization with both the non-negativity and sum-to-one constraint. The trust-constr method is more versatile as it adapts to subdivided problem but has a longer computation time than SLSQP.

2.4) Datasets

No datasets have been specifically produced for this study thus all datasets come from publicly available databases. Mining the literature revealed, interesting datasets for this instance of the deconvolution problem have been published as they contain both single cell and bulk RNA-seq as part of the same experience, other analyzed datasets come from previously published deconvolution method using both bulk and single cell data as input. These potentially interesting datasets are described in Table 1.

Author	**Origin**	**Bulk RNA-seq Accession number**	**Single cell Accession number**
Frishberg et al.	Mus musculus (lungs)	GSE117975	GSE113530
Fadista et al., [29] (bulk) and Baron et al. (single cell)	Human (pancreas)	GSE50244	GSE84133

Table I: Datasets analyzed for this study.

3) Results

3.1) Conceptualized method

The conceptualized method consists in using Seurat to define cell-type specific markers from single cell datasets. Then, deconvolving the related bulk data to obtain estimated cell abundance (Figure 4). An initial objective was to select a method capable of solving the deconvolution problem. It was decided to focus on least squares method. Three method were evaluated: nnls function of SciPy, and both solver of the minimize function from SciPy. One major difference between nnls and minimization is the application of the sum-to-one constraint. For nnls methods the application of this constraint is done after output values were computed (by dividing by the sum). By contrast, sum-to-one constraint is applied during the minimization step when using the SciPy minimize function. Another difference is that we need to give the objective function (function to be minimized) and an initial guess as parameter for SciPy minimize. The objective function is the one allowing to compute our least square residual (Equation 3). The initial guess will be set as the output of SciPy nnls with the same data input. To test those different methods, it was first necessary to obtain and analyze datasets.

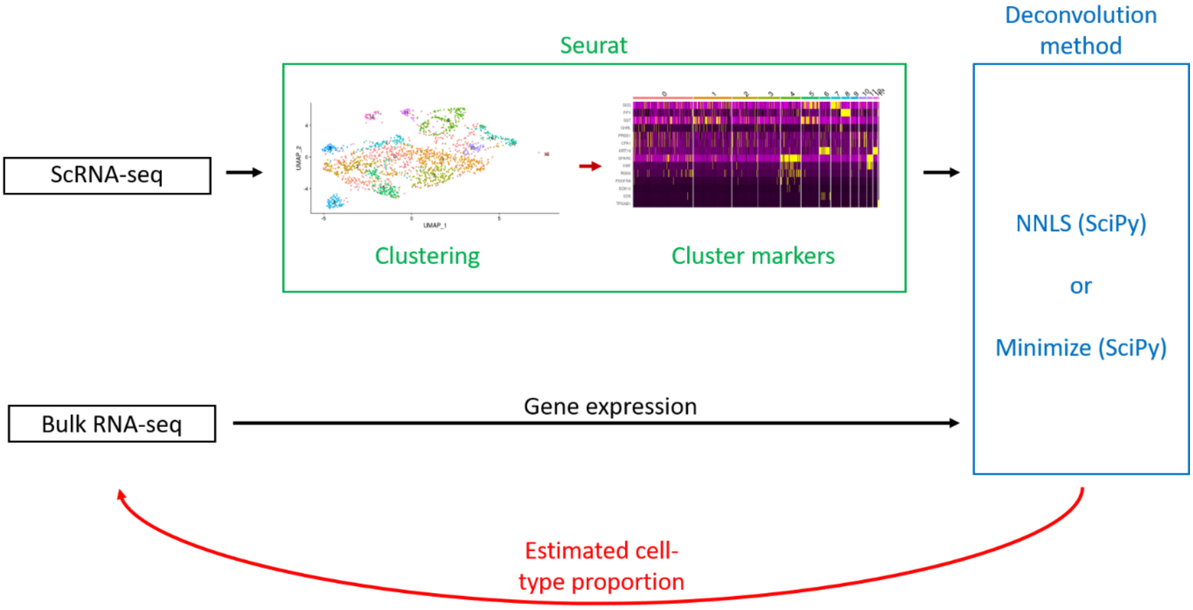


Figure 4: Overview of the conceptualized method.

Seurat is used to infer cluster markers from ScRNA-seq data. Markers and bulk expression are then used as input of the different deconvolution method which estimate cell-type proportion of the bulk sample.

3.2) ScRNA-seq analyses

To be able to set up interesting tests for the deconvolution problem, the first objective is to derive cell specific markers from single cell datasets. Thus, an interesting dataset should feature a heterogeneous cell composition, however the different cell types must be sufficiently different to be separated through clustering and specific markers identified. In that regard, well studied tissues or organs have advantages as previous publications and databases might reference specific markers that would serve as confirmation to the one found through the dataset analysis.

The dataset from Frishberg et al., consists of cells collected from lungs of mouse lines that are specifically bred to simulate human population. The sampling was done after a group of mice was infected with influenza virus while the other served as control. This dataset is interesting as it is an example case were cell composition might act as a confounding factor. Indeed, when differential expression analysis of bulk dataset of healthy versus infected individuals is observed, differences might come from the variation of cell composition in the samples. However, after analyzing this dataset with the Seurat workflow described above, only two clearly defined clusters were obtained (Figure 5). This number of cell types is evidently too low to describe the heterogeneity of lungs tissue. This dataset was thus not used for further analysis.

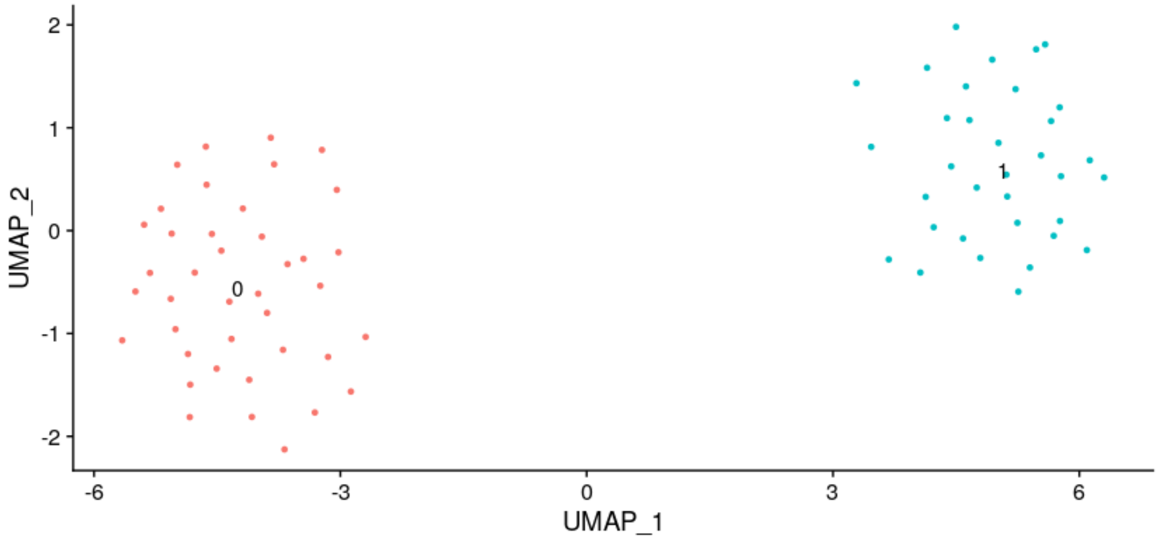


Figure 5: Clusters obtained with Seurat analysis of dataset GSE113530.

The single cell dataset GSE84133 consists of four samples, one for each of the four human donors. There are also two samples from mice that were not further studied as the related bulk dataset only features human samples. The samples were collected from pancreatic islet which is a heterogeneous tissue featuring several different cell types. This dataset was promising as pancreas is a frequently studied organ thus around 15 cell types have been previously characterized with their specific markers. In addition, this dataset has been used to test a deconvolution tool developed by the author which have been able to identify 9 cell types in the first sample and 14 cell types across all the human samples (Figure 6: B, numbers for the other samples are not indicated). However, they used their own method to define cell types and cell markers. So, to gather cell types and related markers from the samples the previously described Seurat workflow was applied to each individual sample. Each sample presented a different number of cell clusters with respectively 12, 13, 11 and 8 clusters (Figure 6: A). However, when using known markers to plot a heatmap some clusters overlapped (Figure 7: A and B) which means that they are probably composed of cells of the same cell type.

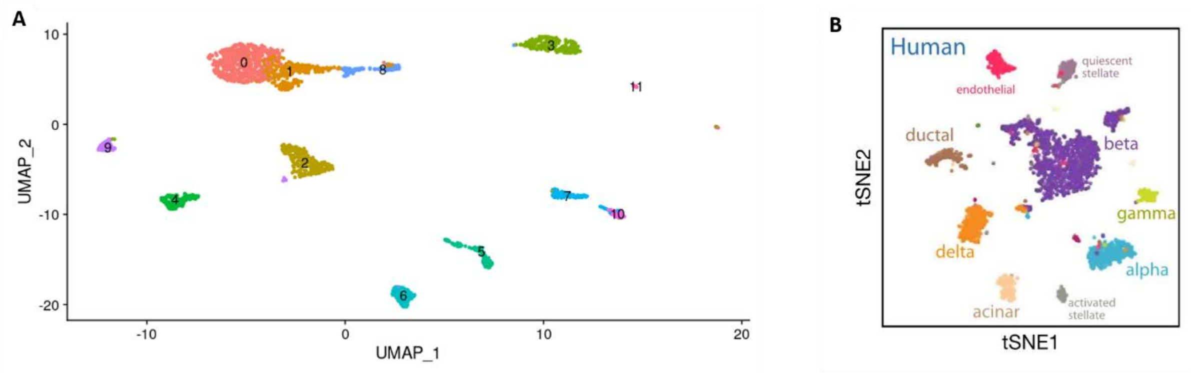


Figure 6: Clustering results of the sample from donor 1.

A: Umap visualization of clusters obtained using Seurat (12 clusters)

B: tSNE visualization of clusters found by the author (9 clusters)

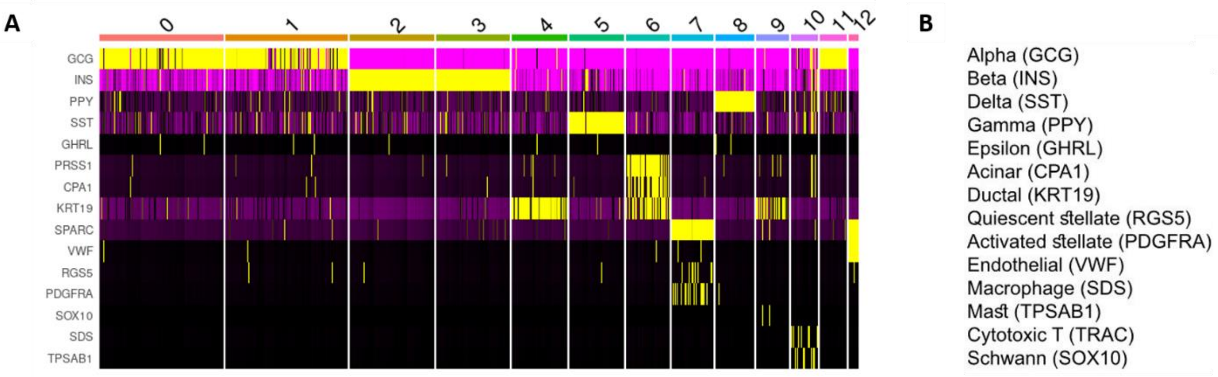
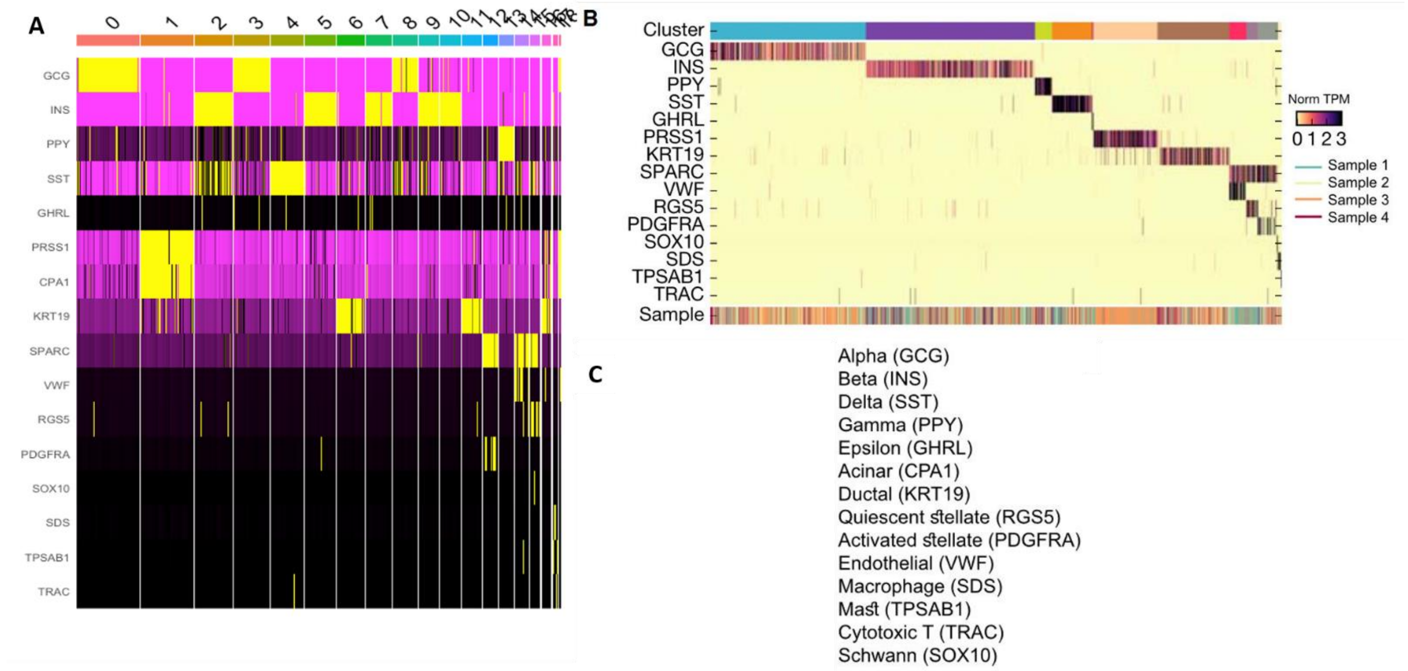


Figure 7: Study of the relation between cluster and cell-type.

A: Heatmap of the known cell-type specific marker in the Seurat clusters for donor 2; **B**: Cell-type and known marker affiliation; The GCG (glucagon) is highly present in cluster 0,1 and 11 which indicates that they are all constituted from Alpha cells, same for clusters 2 and 3 which are probably constituted from Beta cells. Number of identified cell type is then 10.

To obtain the 14 different cell types, the four samples were merged into one which was consecutively analyzed. The merged dataset showed 19 clusters (Figure 9: A, B and C), but the overlap also appeared when plotting the heatmap. This overlap seemed to indicate that the used parameters led to over-clustering. To resolve this over-clustering, the resolution parameter used when computing cluster with Seurat was adjusted. However, modifying the resolution parameter had not the intended effect even if it reduced the number of obtained clusters, it did it by merging non-overlapping clusters (Figure 9: A and B; Figure 10: A and B). In order to keep the smaller non-overlapping clusters, the resolution was kept as it was before. Furthermore, affected clusters were those corresponding to Alpha and Beta cells which have been described by the authors as heterogeneous. Thus, those clusters might represent sub-population.

Once clusters were defined, average expression was calculated for each of their specific markers to represent cell-type specific expression used as input for the deconvolution method



Alpha (GCG)
Beta (INS)
Delta (SST)
Gamma (PPY)
Epsilon (GHRL)
Acinar (CPA1)
Ductal (KRT19)
Quiescent stellate (RGS5)
Activated stellate (PDGFRA)
Endothelial (VWF)
Macrophage (SDS)
Mast (TPSAB1)
Cytotoxic T (TRAC)
Schwann (SOX10)

Figure 8: Overview of cell types identified in the merged sample.

A: Heatmap of the known cell-type specific markers in the Seurat clusters of the merged; **B**: Heatmap of known cell-type specific markers in the author clusters of the merged sample. **C:** Cell-type and known marker affiliation; Alpha and Beta cells are separated in multiple clusters in **A** while they form a single cluster in **B**.

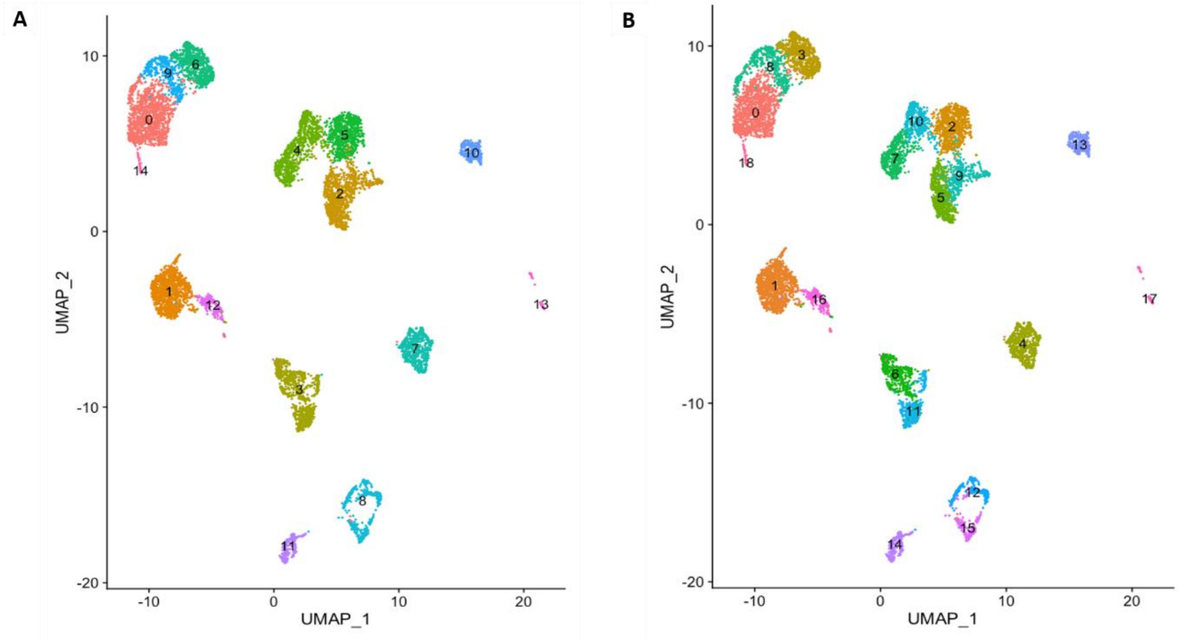


Figure 9: Comparison of Seurat clustering results when modifying the resolution parameter.

A: Umap visualization of Seurat clusters with a resolution set to 0.7; **B**: Umap visualization of Seurat clusters with resolution set to 1 (original value); Lowering the resolution parameter allows to refine cluster borders.

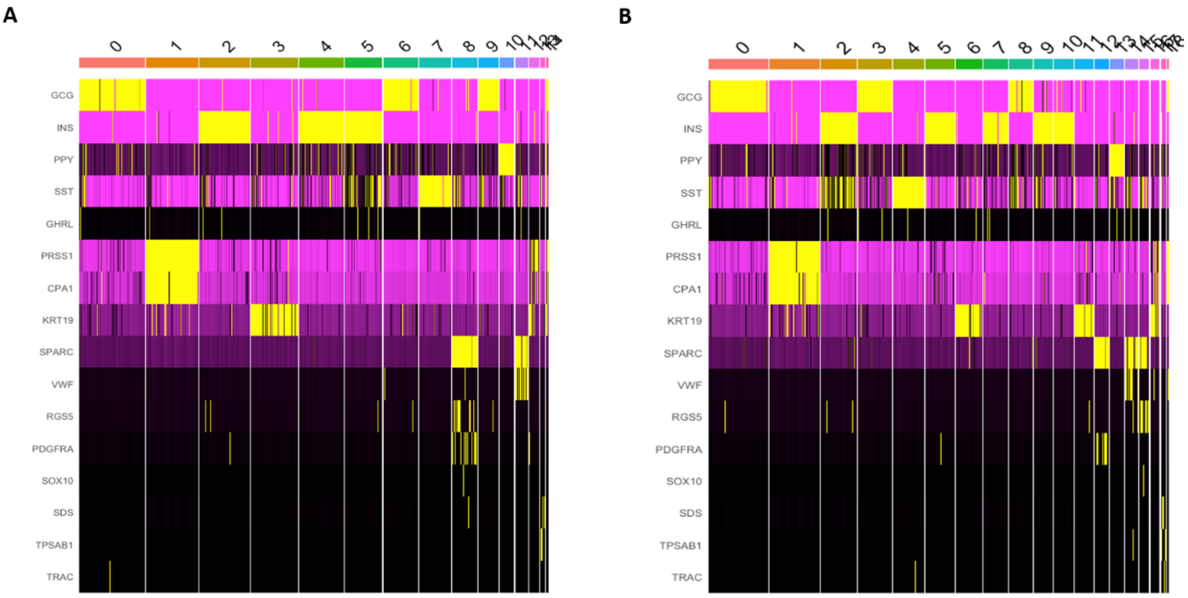


Figure 10: Comparison of Seurat cluster-cell-type affiliation result when modifying the resolution parameter.

A: Heatmap visualization of known markers in Seurat clusters with a resolution set to 0.7; **B**: Heatmap visualization of known markers in Seurat clusters with resolution set to 1 (original value); Lowering the resolution parameter allowed to merge clusters. For example, the 4 Beta cell clusters (INS) in **B** were merged into 3 clusters in **A**. However, a number of small clusters is lost (14 to 19) while some cell types are still separated in multiple clusters (3 Alpha cells cluster (GCG)).

3.2) Deconvolution test

The related bulk dataset from Fadista et al., is composed of RNA sequencing of 89 human pancreatic islet donors. It will serve as the second entry for the deconvolution method. After running the deconvolution method on those 89 samples, we get the estimated proportion of each of the cell types previously defined. Both the SLSQP and Trust-constr solver and the SciPy nnls methods were tested. To study the results, the mean estimated proportion of each cluster was calculated. For the minimize method, while the SLSQP solver is faster (0.03s versus 3s by sample), it gives more extreme results with some cell types overly represented while the others are barely represented. Although the results of the trust-constr method are more spread out, two cell types are given as over abundant by both solvers with proportions respectively around 65% and 27% (Figure 11: B and C). Going back to our cluster markers, we can identify those cell types as being Alpha and Acinar cells. On the other hand, the nnls method gives even more spread out values in even faster computation time. It also estimates that there are two major cell types in the sample with proportions respectively 47% and 29% (Figure 11: A). While clusters are different, they are still affiliated to the same cell types, meaning

Alpha and Acinar cells. However, when observing the most abundant genes, we can see that the Alpha cell type has markers among the highest-ranking ones. As recommended by Mohammedi et al. most expressed genes were filtered. Retesting both solvers and methods without those genes shows much more evenly distributed results and the proportion of previously over-represented Alpha cells dropped considerably while the Acinar cells dropped only slightly.

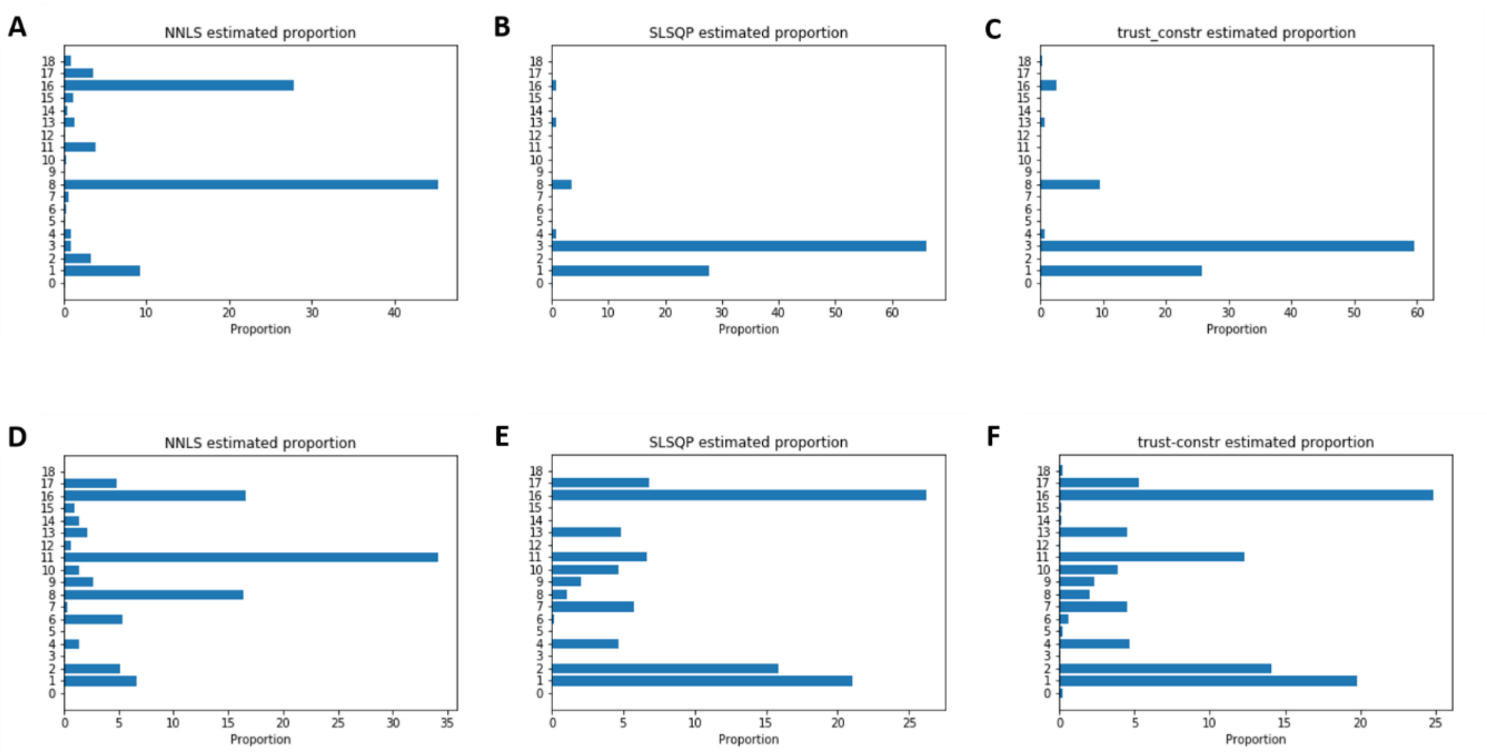


Figure 11: Mean estimated proportion of the 89 samples from Fadista et al., by the different methods.

A, **B** and **C** are estimation from the full marker set. **D**, **E** and **F** are the estimation on the filtered marker set where over abundant genes have been removed.

To be able to assess the quality of the results, simulation tests were set-up by randomly sampling a defined number of cells from the single cell dataset and then adding their expression values to simulate bulk sequencing. Expected proportions are then calculated by considering the cluster assigned to each of the sampled cells. Unfortunately, during the process, the SciPy minimize deconvolution method encountered a bug (probably due to very small values becoming infinite for the computer) and could not finish computing proportions. It was thus not possible to conclude yet concerning the comparison of the different deconvolution methods.

4) Discussion

Because of a lack of time, evaluation of the deconvolution methods was not finished and must be pursued after resolving the encountered issue with the minimize method. Alternatively, other methods capable of addressing the deconvolution problem exist and should also be tested. First, methods such as nnls and minimization from R packages might be better suited as transcriptomic data analysis is frequently done using R. This is also relevant as Seurat which we used to define clusters and markers is a R package. Secondly, including none least square method should give more insight on the most efficient deconvolution method to work on.

Even if the use of scRNA-seq for deconvolution could not be properly evaluated, one key observation is that defining clusters and markers is a central issue on its own. Indeed, ability to define clusters representing cell types is key to be able to use single-cell-referenced deconvolution. Characteristics of good specific markers for cell types should also be well defined as markers were found to have a huge importance in proportion estimation. While the recommendation of Mohammedi et al. to remove over-abundant genes was applied, other deconvolution methods use weight to refine gene influence.

The dataset from Frishberg et al. was probably too small as their deconvolution method was focused on inferring not only cell-type presence but also cell trajectories. To complement this dataset, data from other single cell studies of the same tissue could have been used. Data from Vila Ellis et al. [30] (Table 2) is an example of such dataset. The data from the Ydens et al. [31] publication was analyzed using Seurat but not yet deconvolved as cluster markers were not supported by previously defined cell-type markers. Since having more test situations would increase confidence in the results, other datasets could be analyzed and used to test deconvolution methods such as data from the Sekiguchi et al. [32] publication.

Author	Origin	Bulk RNA-seq Accession number	Single cell Accession number
Vila Ellis et al.,	Mus musculus (lungs)	GSE124324	GSE124323
Ydens et al.	Mus musculus (nervous system)	GSE144705	GSE144707
Sekiguchi et al.	Mus musculus (embryo)	GSE127260	GSE127469

Table II: Potential datasets to complement analysis of this study.

5) Conclusion

While incomplete given the actual circumstances and technical issues, this study represents an initial base to develop a deconvolution method using scRNA-seq derived cell-types for the analysis of bulk RNA-seq data. We were able to conceptualize a model of such a method. Multiple datasets necessary to test this method were identified and a general workflow to analyze them has been defined. First tests were performed on potential deconvolution methods. Those tests allowed to highlight key points of using scRNA-seq such as clustering and specific cell type marker definition.

Those tests must be extended and pursued. They should include more methods and datasets. Recently published deconvolution methods [23]–[26] that similarly use scRNA-seq derived cell type markers should serve as lead to improve method usage such as applying weight to gene expression. They also reveal the extent of potential deconvolution applications to go further than just estimating cell type abundance. In that regard, since previous deconvolution methods used other data than transcriptomic ones, it could be interesting to see if genomic or epigenetic single cell data can also be used to deconvolve related bulk datasets.

On a personal side the study allowed me to gain a better grasp of the process necessary to go from an interesting research topic to an actual research project.

Bibliography

[1] Z. Wang, M. Gerstein, et M. Snyder, « RNA-Seq: a revolutionary tool for transcriptomics », *Nat. Rev. Genet.*, vol. 10, n° 1, p. 57-63, janv. 2009, doi: 10.1038/nrg2484.

[2] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, et Y. Gilad, « RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays », *Genome Res.*, vol. 18, n° 9, p. 1509-1517, janv. 2008, doi: 10.1101/gr.079558.108.

[3] K. R. Kukurba et S. B. Montgomery, « RNA Sequencing and Analysis », *Cold Spring Harb. Protoc.*, vol. 2015, n° 11, p. pdb.top084970, janv. 2015, doi: 10.1101/pdb.top084970.

[4] A. Conesa *et al.*, « A survey of best practices for RNA-seq data analysis », *Genome Biol.*, vol. 17, n° 1, p. 13, janv. 2016, doi: 10.1186/s13059-016-0881-8.

[5] A. Kuhn, A. Kumar, A. Beilina, A. Dillman, M. R. Cookson, et A. B. Singleton, « Cell population-specific expression analysis of human cerebellum », *BMC Genomics*, vol. 13, n° 1, p. 610, nov. 2012, doi: 10.1186/1471-2164-13-610.

[6] R. Bacher et C. Kendziorski, « Design and computational analysis of single-cell RNA-sequencing experiments », *Genome Biol.*, vol. 17, p. 63, avr. 2016, doi: 10.1186/s13059-016-0927-y.

[7] A. Haque, J. Engel, S. A. Teichmann, et T. Lönnberg, « A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications », *Genome Med.*, vol. 9, n° 1, p. 75, août 2017, doi: 10.1186/s13073-017-0467-4.

[8] B. Hwang, J. H. Lee, et D. Bang, « Single-cell RNA sequencing technologies and bioinformatics pipelines », *Exp. Mol. Med.*, vol. 50, n° 8, p. 96, 07 2018, doi: 10.1038/s12276-018-0071-8.

[9] B. Vieth, S. Parekh, C. Ziegenhain, W. Enard, et I. Hellmann, « A systematic evaluation of single cell RNA-seq analysis pipelines », *Nat. Commun.*, vol. 10, n° 1, p. 4667, oct. 2019, doi: 10.1038/s41467-019-12266-7.

[10] G. Chen, B. Ning, et T. Shi, « Single-Cell RNA-Seq Technologies and Related Computational Data Analysis », *Front. Genet.*, vol. 10, avr. 2019, doi: 10.3389/fgene.2019.00317.

[11] O. B. Poirion, X. Zhu, T. Ching, et L. Garmire, « Single-Cell Transcriptomics Bioinformatics and Computational Challenges », *Front. Genet.*, vol. 7, 2016, doi: 10.3389/fgene.2016.00163.

[12] B. Ding *et al.*, « Normalization and noise reduction for single cell RNA-seq experiments », *Bioinformatics*, vol. 31, n° 13, p. 2225-2227, juill. 2015, doi: 10.1093/bioinformatics/btv122.

[13] P. Brennecke *et al.*, « Accounting for technical noise in single-cell RNA-seq experiments », *Nat. Methods*, vol. 10, n° 11, p. 1093-1095, nov. 2013, doi: 10.1038/nmeth.2645.

[14] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, et J. C. Marioni, « Normalizing single-cell RNA sequencing data: challenges and opportunities », *Nat. Methods*, vol. 14, n° 6, p. 565-571, juin 2017, doi: 10.1038/nmeth.4292.

[15] F. Avila Cobos, J. Vandesompele, P. Mestdagh, et K. De Preter, « Computational deconvolution of transcriptomics data from mixed cell populations », *Bioinformatics*, vol. 34, n° 11, p. 1969-1979, juin 2018, doi: 10.1093/bioinformatics/bty019.

[16] X. Chen, S. A. Teichmann, et K. B. Meyer, « From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture », *Annu. Rev. Biomed. Data Sci.*, vol. 1, n° 1, p. 29-51, 2018, doi: 10.1146/annurev-biodatasci-080917-013452.

[17] T. Erkkilä, S. Lehmusvaara, P. Ruusuvuori, T. Visakorpi, I. Shmulevich, et H. Lähdesmäki, « Probabilistic analysis of gene expression measurements from heterogeneous tissues », *Bioinformatics*, vol. 26, n° 20, p. 2571-2577, oct. 2010, doi: 10.1093/bioinformatics/btq406.

[18] S. Mohammadi, N. Zuckerman, A. Goldsmith, et A. Grama, « A Critical Survey of Deconvolution Methods for Separating cell-types in Complex Tissues », *Proc. IEEE*, vol. 105, n° 2, p. 340-366, févr. 2017, doi: 10.1109/JPROC.2016.2607121.

[19] S. S. Shen-Orr et R. Gaujoux, « Computational deconvolution: extracting cell type-specific information from heterogeneous samples », *Curr. Opin. Immunol.*, vol. 25, n° 5, p. 571-578, oct. 2013, doi: 10.1016/j.coi.2013.09.015.

[20] V. K. Yadav et S. De, « An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples », *Brief. Bioinform.*, vol. 16, n° 2, p. 232-241, mars 2015, doi: 10.1093/bib/bbu002.

[21] D. Venet, F. Pecasse, C. Maenhaut, et H. Bersini, « Separation of samples into their constituents using gene expression data », *Bioinformatics*, vol. 17, n° Suppl 1, p. S279-S287, juin 2001, doi: 10.1093/bioinformatics/17.suppl_1.S279.

[22] Z. Li, Z. Guo, Y. Cheng, P. Jin, et H. Wu, « Robust partial reference-free cell composition estimation from tissue expression », *Bioinformatics*, doi: 10.1093/bioinformatics/btaa184.

[23] A. Frishberg *et al.*, « Cell composition analysis of bulk genomics using single-cell data », *Nat. Methods*, vol. 16, n° 4, p. 327-332, avr. 2019, doi: 10.1038/s41592-019-0355-5.

[24] M. Baron *et al.*, « A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure », *Cell Syst.*, vol. 3, n° 4, p. 346-360.e4, oct. 2016, doi: 10.1016/j.cels.2016.08.011.

[25] X. Wang, J. Park, K. Susztak, N. R. Zhang, et M. Li, « Bulk tissue cell type deconvolution with multi-subject single-cell expression reference », *Nat. Commun.*, vol. 10, n° 1, p. 1-9, janv. 2019, doi: 10.1038/s41467-018-08023-x.

[26] M. Dong *et al.*, « SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references », *Brief. Bioinform.*, doi: 10.1093/bib/bbz166.

[27] R Core Team, « R: A Language and Environment for Statistical Computing ». R Foundation for Statistical Computing, 2020, [En ligne]. Disponible sur: https://www.R-project.org.

[28] P. Virtanen *et al.*, « SciPy 1.0: fundamental algorithms for scientific computing in Python », *Nat. Methods*, vol. 17, n° 3, p. 261-272, mars 2020, doi: 10.1038/s41592-019-0686-2.

[29] J. Fadista *et al.*, « Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism », *Proc. Natl. Acad. Sci.*, vol. 111, n° 38, p. 13924-13929, sept. 2014, doi: 10.1073/pnas.1402665111.

[30] L. Vila Ellis *et al.*, « Epithelial Vegfa Specifies a Distinct Endothelial Population in the Mouse Lung », *Dev. Cell*, vol. 52, n° 5, p. 617-630.e6, mars 2020, doi: 10.1016/j.devcel.2020.01.009.

[31] E. Ydens *et al.*, « Profiling peripheral nerve macrophages reveals two macrophage subsets with distinct localization, transcriptome and response to injury », *Nat. Neurosci.*, vol. 23, n° 5, p. 676-689, mai 2020, doi: 10.1038/s41593-020-0618-6.

[32] R. Sekiguchi, D. Martin, et K. M. Yamada, « Single-Cell RNA-seq Identifies Cell Diversity in Embryonic Salivary Glands », *J. Dent. Res.*, vol. 99, n° 1, p. 69-78, janv. 2020, doi: 10.1177/0022034519883888.