Universiteit Leiden

Master Computer Science

Comparison of expression deconvolution methods

Name:	Daniël Wijnbergen
Student ID:	s2376067
Date:	30/06/2020
Specialisation:	Bioinformatics
1st supervisor:	Dr. K.J. Wolstencroft, LIACS
2nd supervisor:	Prof. Dr. D.J.M. Peters, LUMC

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Polycystic kidney disease is a disease occurring in 1 in 1000 live births. It is usually caused by a mutation in the PKD1 or PKD2 gene. The mutation causes the formation of cysts which as the disease progresses, eventually leads to renal failure. PKD is a disease in which multiple cell types including epithelial cells, macrophages and fibroblasts are involved.

In a previous project, MuSiC was used to estimated the proportions of various cell types in both healthy and diseased kidney tissue. The results of MuSiC were missing some of the cell types and it was not clear how reliable the results were. Because other methods exist that could offer a more accurate deconvolution, and to get more insight in the reliability of deconvolution on a complex dataset, six methods including MuSiC were benchmarked on one simulated dataset and one experimental bone marrow dataset.

MuSiC, DWLS and CIBERSORTx without batch correction performed the best on the simulated dataset while DWLS and CIBERSORTx with batch correction performed the best on the bone marrow dataset. DWLS seems to be the most accurate method while CIBERSORTx was nearly as accurate but was much faster and used far less memory than DWLS. Although some patterns in proportion changes were correctly estimated, other estimations were not as accurate.

MuSiC, DWLS and CIBERSORTx were also executed on the PKD dataset and estimated an increase in macrophages and fibroblasts in samples with the severe disease phenotype. This confirmed the earlier results of MuSiC for the macrophages and is in agreement with the involvement of macrophages and fibroblasts in the disease. The decrease in proximal tubular cells estimated by MuSiC was not seen as strongly in DWLS and CIBERSORTx.

Finally, the estimated proportions in PKD were multiplied with cell type gene expression in order to estimated the gene expression changes solely as a result of changes in cell type proportions. For example, a macrophage marker gene was estimated to have a five-fold increase in expression, because the proportion of macrophages was increased in the samples with a severe PKD phenotype.

Contents

1 Introduction

1.1 Polycystic Kidney Disease

Polycystic Kidney Disease (PKD) is a genetic disease occurring in 1 in 1000 live births [1]. PKD causes the formation of cysts in the kidneys. The renal function of the kidney's decrease as the cysts increase in size, eventually causing renal failure.

PKD is caused by a mutation in either the PKD1 or PKD2 gene with mutations in PKD1 causing a more severe disease phenotype. PKD1 and PKD2 encode for the polycystin 1 and polycystin 2 proteins respectively. These proteins form a complex that mostly functions in the primary cilia. Although the complex is likely to be involved in processes like mechanosensitivity and Wnt/Ca2+ signaling [2, 3], the exact function is not clear.

One factor in PKD is the involvement of various cell types in the disease. The cysts themselves are formed by epithelial cells, but some other cell types including fibroblasts and macrophages are also involved in PKD after kidney injury [4]. For example, Macrophages promote cyst growth in PKD [5, 6]. Similarly, Fibroblast Growth Factor 23 was associated with kidney function decline in PKD [7]. Another important change in the cell types during PKD is the dedifferentiation of epithelial cells. Up-regulated genes during the disease were found to be in gene sets related to renal development [8]. During renal injury, these cells differentiate and proliferate in order to repair kidney injury.

1.2 Transcriptomics

One of the methods used for research into PKD is the measurement of the gene transcripts in kidney tissue in order to analyse global changes in gene expression in the disease state. Because these transcripts are translated into proteins, the number of gene transcripts in a cell, roughly indicates the number of proteins in a cell. The measurement of the number of transcripts is currently usually done using RNA-Sequencing. Transcripts are isolated, fragmented and then reverse-transcribed into cDNA. Each fragment is then sequenced in order to determine the sequence of each individual fragment. Finally, the obtained sequences are compared to reference sequence in order to count to number of transcripts originating from each gene.

Previously, RNA-Sequencing was performed on PKD tissues. Multiple signaling pathways that are involved in the PKD were identified [9]. However, a limitation of most RNA-Sequencing experiments including this one is that the sequencing is done for an entire tissue. This means that the transcripts of multiple cell types are all mixed together in one sample and it is unclear how the gene expression changes relate to the cell types. This is especially relevant in PKD where multiple cell types and the signaling between them are involved.

A relatively recent development in transcriptomics that addresses this limitation is single cell RNA sequencing [10]. The goal of scRNA sequencing is to obtain the transcripts counts for individual cells instead of the whole sample or

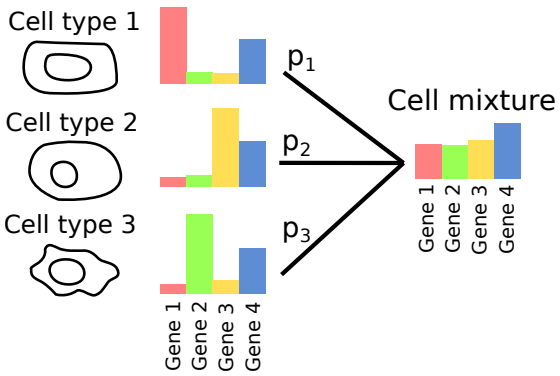


Figure 1: The basic idea of deconvolution. The transcripts of cell types in a tissue are mixed with proportions p_1, p_2 and p_3. Deconvolution methods estimate these proportions from the cell type specific data and the cell mixture data.

tissue. One method to achieve this is to add a unique barcode sequence to the transcripts each cell. During the analysis of the raw sequencing data, these barcodes are then used to determine the originating cell of each transcript. scRNA sequencing comes with some downsides however. The number of reads is lower because it is spread over all cells, the cost of the protocol is higher and the dissociation of the cells in a tissue causes a bias towards easily dissociated cell types and towards stress response genes [11].

1.3 Expression deconvolution

Many approaches have been developed that estimate the proportions of various cell types in the transcriptomics of a tissue sample. This problem is usually called expression deconvolution. The basic idea of deconvolution is shown in figure 1. In a cell mixture or tissue, the transcripts of different cell types are mixed with certain proportions (in this case p_1, p_2 and p_3). The goal of deconvolution is then to estimate these proportions from the gene expression measurements from both single cell types and the cell mixture or tissue. In the case of the example, the proportions of all three cells are roughly $\frac{1}{3}$. Adding the expression values of the three cell types with equal weight roughly results in the expression value in the cell mixture.

In an earlier project, MuSiC [12] was applied on bulkRNA data for PKD where it estimated an increase in macrophages and a decrease in proximal tubular cells as the disease severity increases. However, some of the cell types were not detected at all in the bulkRNA data. Furthermore, it was not clear how reliable MuSiC is when applied on a very complex dataset and if other expression deconvolution methods exist that perform better than MuSiC. For these reasons, MuSiC and multiple other expression deconvolution methods will be benchmarked on two datasets.

Many approaches have been developed for the problem of expression deconvolution. The focus of the benchmark will be on methods that use scRNA data as a reference. The 6 methods that are compared are: MuSiC [12], SCDC [13], MOMF [14], CIBERSORTx [15], DigitalDLSorter [16] and DWLS [17]. Each of these methods function slightly differently. For example, MuSiC is based on weighted linear regression which focuses on informative genes that do not vary much between samples, CIBSERSORTx is based on support vector regression which is robust to noise in the gene expression and DigitalDLSorter is based on deep learning.

1.4 Goals

The main research questions are:

- Which scRNA based expression deconvolution methods perform the best on a complex dataset?
- What are the changes in cell proportions estimated by these methods in PKD?

2 Methods

The methods were benchmarked on different 2 datasets. The first one is a simulated dataset and the second one is an experimental bone marrow dataset [18]. Each method was run with the default parameters except for the batch corrected version of CIBERSORTx. CIBERSORTx was run with S-batch correction and 0 as the fraction parameter. The batch corrected version of CIBERSORTx is referred to as CIBERSORTxCorrected in the results. The resulting proportions of all methods were compared to the "true" proportions in order to quantify the performance of the methods. The pearson correlation and Root Mean Square Error (RMSE) were used as the metrics because they were the most used metrics in research on expression deconvolution. The best performing methods according to these metrics were also applied on kidney data [9, 19].

2.1 Workflow

The workflow for the benchmarks is shown in figure 2. First, the simulated or experimental datasets are prepared for the benchmark, which uses three csv files. One with the single cell expression values, one with the bulk expression values and one with the true cell type proportions. The single cell and bulk expression data are then used as the input for each deconvolution method separately. For each deconvolution method, a script turns the input data into a format that the specific methods needs. The predicted proportions for the cell types are then written to a csv file for that method. These proportion files are then compared to the true proportion matrix from the data. All scripts are available at `git.lumc.nl/dwijnbergen/deconvolution-benchmark`.

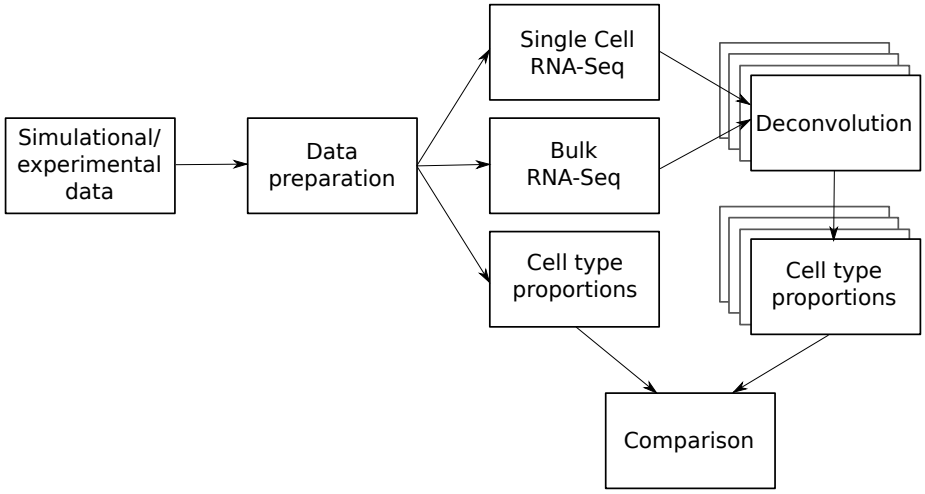


Figure 2: Workflow of the benchmarks. The input dataset is transformed into a scRNA, bulkRNA, and proportion file. These files are then used as the input for the deconvolution methods. The results of the deconvolution methods are finally compared to the proportions from the dataset.

2.1.1 Data formats

Each type of data is stored as a csv file without quotation marks. For the scRNA data, each row is a gene and each column is a cell with their identifiers stored in the first row and column. Each value in the file is the number of counts for that combination of cell and gene. The exceptions here are the first two rows which are the cell type name and subject id. The bulkRNA data is similar except the columns are samples instead of cells. Finally, in the proportions files, the samples are the columns and the cell types are the rows, each value is then the proportion of that cell type in that sample.

2.2 Deconvolution

Mathematically, the deconvolution can be defined as

$$\mathbf{Sp} = \mathbf{b}$$

Where $\mathbf{S}$ is the $n \times k$ cell type signature matrix, $\mathbf{p}$ is the $k \times 1$ proportion vector and $\mathbf{b}$ is the $n \times 1$ bulk expression vector (n is the number of genes and k is the number of cell types) [17]. The values of vector $\mathbf{p}$ are then determined using $\mathbf{S}$ and $\mathbf{b}$. Because there is no exact solution, most methods rely on minimizing various error metrics of this equation. Each method is described shortly in this section.

2.2.1 MuSiC

MuSiC [12] is a weighted non-negative least squares (W-NNLS) regression based method. This means that the predicted proportions are changed in order to optimize the least squares error between the bulk RNA-Seq expression levels and the scRNA-Seq expression levels multiplied by the proportions of the cell types. Additionally, the proportions of the cell types are constrained to be non-negative since this is not possible in reality.

Additionally, each gene is given a weight based on how informative the gene is. This weight is based on both how much the gene varies between cell types and how much the gene varies between subjects. The weight is higher if it varies between cell types and is lower when it varies between subjects. The reasoning behind the subject-variance based weighting is that genes that are similar in different subjects for the scRNA data will likely also be similar for the subjects in the bulkRNA data.

2.2.2 SCDC

SCDC [13] is very similar to MuSiC. It also uses a W-NNLS framework, but differs in the calculation of the weights. Furthermore, SCDC is able to use an ensemble where multiple different scRNA-Seq data sources can be used in parallel. The weighted average of each scRNA-Seq dataset based proportions are then used as the predicted proportion. scRNA-Seq data sources that result in a regression with a better fit are given a higher weight in the averaged proportions.

2.2.3 MOMF

In MOMF [14], the deconvolution problem is solved as a Non-negative Matrix Factorization (NMF) problem. This is a class of algorithms where a matrix is factorized into two matrices and all matrices involved only have positive values.

The main difference between MOMF and other methods is that it models the count nature of gene expression data. This means that the method accounts for the uncertainty in the bulkRNA-Seq, RNA-Seq gene expression levels and also the cell type specific mean gene expression levels. This is done by instead of directly using the count data to execute the deconvolution on, the "true" expression is a variable in the model. In the model, the "true" expression should be as close together to the count data as possible but should also result in a deconvolution with low residuals.

2.2.4 CIBERSORTx

CIBERSORTx [15] is based on ν-support vector regression (ν-SVR). This is a type of regression that is resistant to noise. In regular regression, the deviation of the fitted line to the is minimized. Support vector regression works similarly, but there is a margin around the fitted line where the error is calculated from instead of the line itself. Small deviations from the fitted line (that are inside of the margin) are therefore ignored. In ν-SVR, the number of datapoints that

are allowed to be outside of the margin (Support vectors) is controlled by the parameter *nu*. The margin is then adjusted by in order to have the correct number of data point outside of the margin.

Additionally, CIBERSORTx includes two batch correction methods that attempt to minimize the batch effect between the bulk and single cell data in order to minimize the deconvolution error.

2.2.5 DigitalDLSorter

DigitalDLSorter [16] is a machine learning method utilising a neural network. The method first adds simulated cells to the scRNA data based on parameters estimated from the scRNA existing data. After that, a large number of simulated bulk samples with known proportions are created from the scRNA data. These bulk samples with known proportions are used as the training samples for a neural network where the gene expression data of individual bulkRNA samples are used as the input for the network while the proportions of all cell types are used as the expected output of the network. In the network, there are two hidden layers with 200 neurons between the input and output. Finally, the trained neural network can be applied on the real bulk data in order to predict their cell type proportions.

2.2.6 DWLS

DWLS [17] (Dampened Weighted Least squares) is a regression method that is focused on the correct estimation of rare cell types. DWLS aims to solve two problems inherent to the least squares regression model. The first is the high estimation error of rare cell types caused by the low impact of these on the least squares error. The second is the low contribution of informative genes due to their low mean expression values.

DWLS attempt to solve these problems by giving genes that have a high expression (derived from the cell type signature matrix and the estimated proportion vector) a lower weight. Because the weight are calculated using the estimated proportions, DWLS first estimated the proportions without weighting and then iteratively estimates new proportions using the weights from the last iteration.

2.2.7 Ensemble

An ensemble of these six methods was also benchmarked alongside the methods themselves. This ensemble works by calculating the average proportion estimations by all other methods. This is similar to the ensemble in SCDC [13], but unweighted instead of weighted and applied on multiple methods instead of multiple scRNA datasets.

2.3 Simulated data

A simple method was used to generate fully controllable generated expression data and cell type proportions. The overview of this method is shown in figure 3. The method consists of four steps.

2.3.1 Cell type gene expression profiles

First, the gene expression profiles for all cell types are generated. The expression of each gene is determined by taking the value 2 squared by a sample from a uniform distribution from 0 to 10 ($X \sim 2^{U(0,10)}$). Differentially expressed genes between cell types are then simulated by re-sampling all values of these genes with a probability of 0.2.

2.3.2 scRNA expression

Each cell in the scRNA data is generating by copying the cell type gene expression profile for the relevant cell type and applying some steps on them. First, in order to simulate biological variation, each value is replaced with a sampled value from a normal distribution where μ is the original value and σ is 10% of the original value. Secondly, because scRNA data is discrete and has a low read depth, the value is multiplied by 0.05 and then rounded. Finally, in order to simulate the variation caused by sequencing, each value is replaced with a sampled value from the Poisson distribution where λ is the original value.

2.3.3 bulkRNA Proportions

The proportion of each cell type in all samples is first sampled from an uniform distribution from 1 to 20. Each value is then multiplied with N(1, 0.1) in order to create variation of the proportions between the samples. Finally, the values are divided by the total proportions for their sample in order to make the sum of the proportions 1 for each sample. It is worth noting that all cell types in the simulated bulkRNA data are defined in the scRNA data, which is not necessarily true for real data.

2.3.4 bulkRNA expression

The bulkRNA expression is generated by multiplying the cell type gene expression profile matrix with the proportion matrix. The values are rounded and then used as the λ in the Poisson distribution. The samples from the Poisson distribution are then the final expression values for the bulkRNA data.

2.4 Bone marrow

In addition to the simulated dataset, the methods were also tested with a real dataset. This dataset is a dataset with scRNA-Seq, bulk RNA-Seq and flow cytometry data from bone marrow [18]. Because these three types of data are

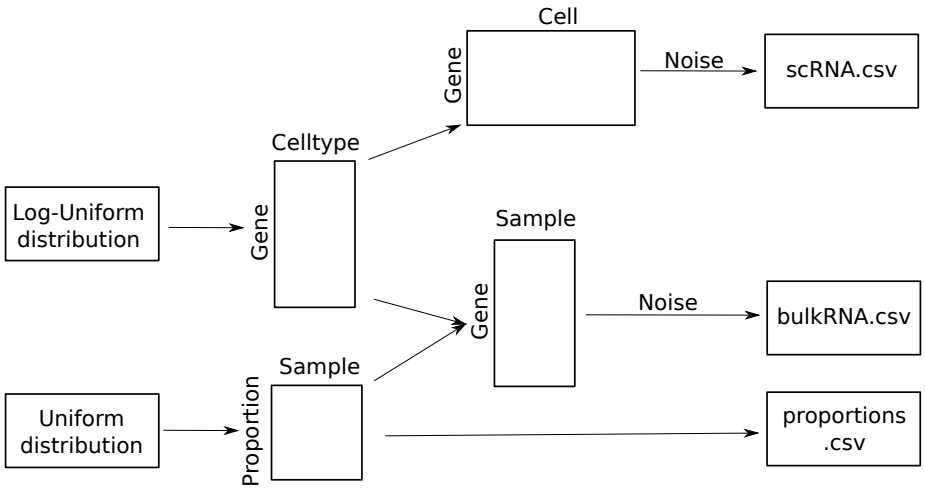


Figure 3: Overview of the simulation method. Matrices for cell type expression profiles and cell type proportions are generated from random distributions. These matrices are then used to generate scRNA and the bulkRNA matrices.

collected from the same samples, the deconvolution can be applied on one of the datasets while the measured proportions for the same samples can be obtained from another dataset. The datasets are generated from bone marrow samples of 20 healthy volunteers and contain mostly immune system related cell types.

The bulkRNA and scRNA data were downloaded from the Gene Expression Omnibus (GEO) using accession code GSE120446. The file with the cell type annotations was requested and received from the authors. For the bulkRNA data, the RNASeqCounts file was downloaded and transformed into the input format for the benchmark scripts. For the scRNA data, each single cell matrix for the individuals in the study were combined. For each cell type, a maximum of 1000 cells were sampled in order to limit the amount of prior information about the cell type proportions that is able to leak into the deconvolution methods and simultaneously to reduce the computational cost of the deconvolution. Because the scRNA proportions were close to the flow cytometry proportions, the proportions in the scRNA data were used as the "real" proportions for the bulk data from the same sample.

2.5 PKD

Finally, the deconvolution methods that performed the best in the benchmarks were applied on a bulkRNA dataset for PKD [9]. In this dataset, there are 3 mice with a severe PKD phenotype, 4 mice with an intermediate phenotype and 5 with a wild-type phenotype. The PKD phenotype in the mice is caused by a kidney-specific inducible knockout in the PKD1 gene. This knockout is induced 38 days after the birth of the mice. The gene expression was measured

at 18 and 21 weeks after birth for intermediate and severe samples respectively. A csv file with the expression data and expression set with the metadata were extracted from the excel file. The gene identifiers were mapped to their gene symbols using BioMart in order to be useable with scRNA datasets that use gene symbols [20].

For the corresponding kidney scRNA dataset, the dataset by Park et al. was used [19]. This dataset contains the scRNA data obtained from 7 mice, of which 4 are wild-type and 3 have various lineage tags. The scRNA data was clustered by Park et al. into 16 different cell types including epithelial and immune cell types. The data was downloaded from the Gene Expression Omnibus with accession code GSE107585 and then transformed into format used in the benchmarks. The two novel cell types were removed from the dataset.

2.6 Convolution

The estimated proportions in PKD were multiplied with wild type gene expression data from different cell types (average per cell type from scRNA data) in order to estimate the expected fold changes in bulkRNA data from gene as a result solely of the changes in proportions (figure 4). For the estimation of fold changes with all proportion changes combined, the average proportions of the wild type bulk and the average proportions of severe PKD bulk samples were multiplied with the cell type gene expression profiles and then compared. These estimated fold changes were then written to the first column of a csv file.

In order to see which cell type proportion change explains the expected fold change, proportion vectors in which only one cell type is changed to the proportion in severe PKD were also multiplied with the cell type gene expression. The rest of the cells were multiplied with a value in order to keep the sum of all proportions at 1. The resulting vector of gene expression values could then be compared with the wild type vector in order to calculate the estimated changes in gene expression as a result of only the change in proportion of that cell type. These fold changes were written to a column in the csv file corresponding to the changed cell type. An example of the convolution is shown in appendix A.

3 Results

3.1 Simulation

The six deconvolution methods were first benchmarked on a simulated dataset. In this simulation, cell type specific gene expression profiles and bulkRNA cell type proportions are generated from random distributions and then used to generate scRNA and bulkRNA data. The following parameters were used for the simulation:

- 20 cell types
- 5000 genes

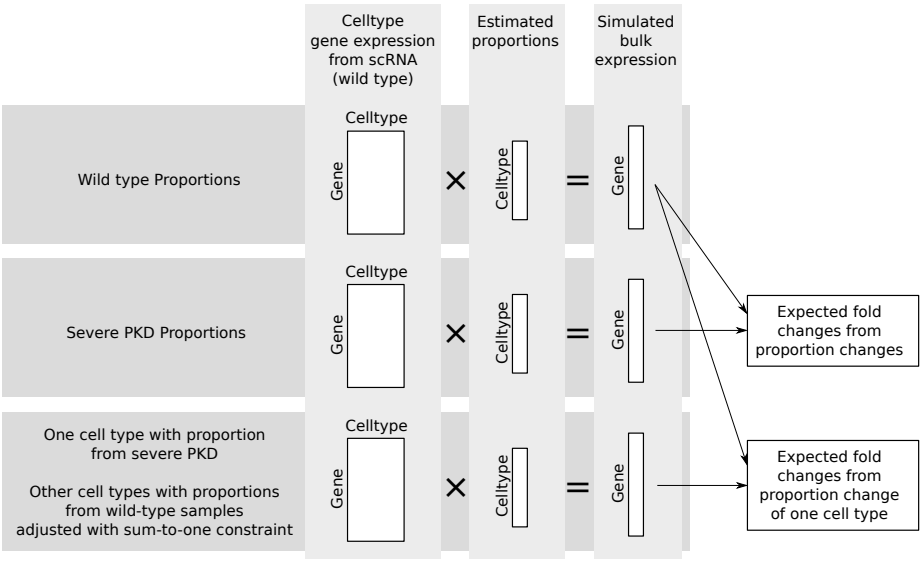
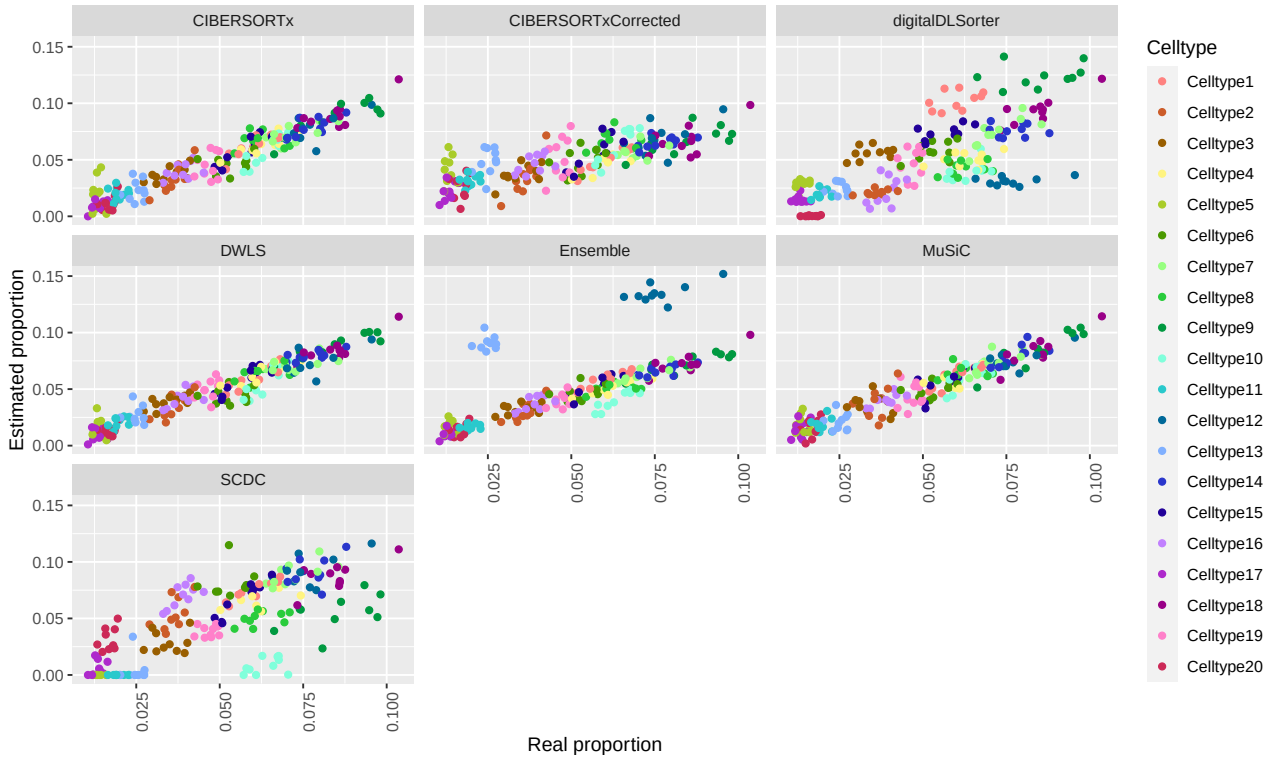


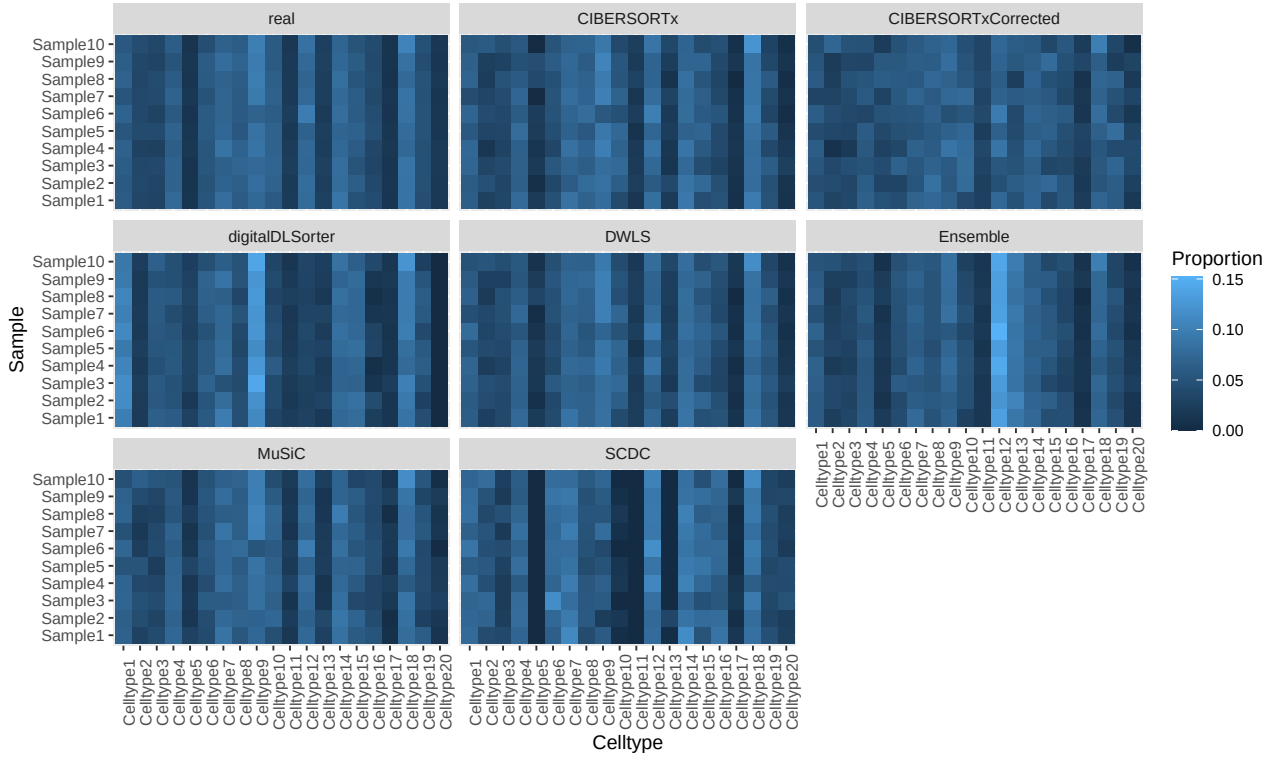
Figure 4: Three different types of proportion vectors were multiplied with the cell type specific gene expression matrix calculated from the scRNA data. The resulting gene expression vectors are then compared to each other to show expected fold changes in gene expression as a result of cell type proportion changes.

- 200 differentially expressed genes
- 10 bulkRNA samples
- 3 subject (subject variation not simulated)
- 100 cells per cell type

All methods were applied on the datasets with default parameters except for the batch corrected CIBERSORTx (CIBERSORTxCorrected) which was used with the S-batch correction mode turned on and the fraction parameter on 0. The results of the benchmark on the simulated data are shown in figure 5 without MOMF and in appendix B with MOMF. The Pearson correlation and RMSE for each method are shown in table 1. The best results on the simulated data were achieved by CIBERSORTx, DWLS and MuSiC which all have a Pearson correlation above 0.95. There is no clear difference in the performance for these methods. SCDC and digitalDLSorter had a few cell types that were over or underestimated (figure 5). Usually, if the proportions of a cell type was overestimated in one simulated bulkRNA sample, it was also overestimated for the other simulated bulkRNA samples. For example, all cells of cell type 1 are overestimated. MOMF incorrectly predicted extremely high proportions for cell types 12 and 13 while predicting a proportion of 0 for the other cell types.



(a) Scatter plot



(b) Heatmap

Figure 5: Benchmark results on simulated data

Table 1: The Pearson correlation and root mean square error (RMSE) of the methods for the simulation data.

Method	Pearson correlation	RMSE
CIBERSORTx	0.951	0.0083
digitalDLSorter	0.745	0.0217
DWLS	0.959	0.0071
Ensemble	0.680	0.0220
MOMF	0.025	0.1514
MuSiC	0.951	0.0077
SCDC	0.747	0.0224

Table 2: The Pearson correlation and root mean square error (RMSE) of the methods for the bone marrow data.

Method	Pearson correlation	RMSE
CIBERSORTx	0.267	0.0886
CIBERSORTxCorrected	0.506	0.0569
DWLS	0.587	0.0742
Ensemble	0.309	0.0645
MOMF	0.078	0.0925
MuSiC	-0.110	0.1192
SCDC	-0.016	0.1259

3.2 Bone marrow

The Bone Marrow dataset was first validated to see if the scRNA and bulkRNA values are correlated. The resulting comparison (for genes with more than 1 count after log2 scaling) is shown in figure 6. There is a clear linear relation between the gene expression in the bulkRNA data and the scRNA data after log2 scaling. The Pearson correlation for the log2 scaled values is 0.84 with a p-value below 2.2^{-16}. Despite the linear correlation, there are still some very large differences between the scRNA and bulk data. Some genes have a far higher expression in one of that datasets than the other. This is likely to decrease the accuracy of the estimated proportions.

The bulkRNA dataset was subsequently deconvoluted using the scRNA dataset. The results of the deconvolution are shown in figure 7. The Pearson correlation and RMSE for each method are shown in table 2. The best results according to Pearson correlation were DWLS, CIBERSORTx with correction and then CIBERSORTx without correction in descending order. The Pearson correlations were 0.587, 0.506 and 0.267 respectively. Interestingly, CIBERSORTx with correction had a lower RMSE than DWLS. MOMF, MuSiC and SCDC did not perform well and got a Pearson correlation close to zero.

The computational performance of the methods is shown in table 3. MuSiC

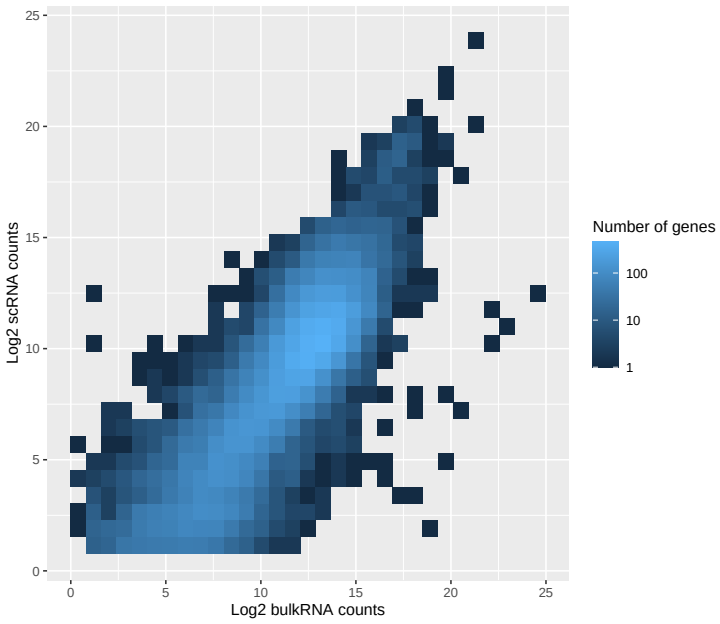
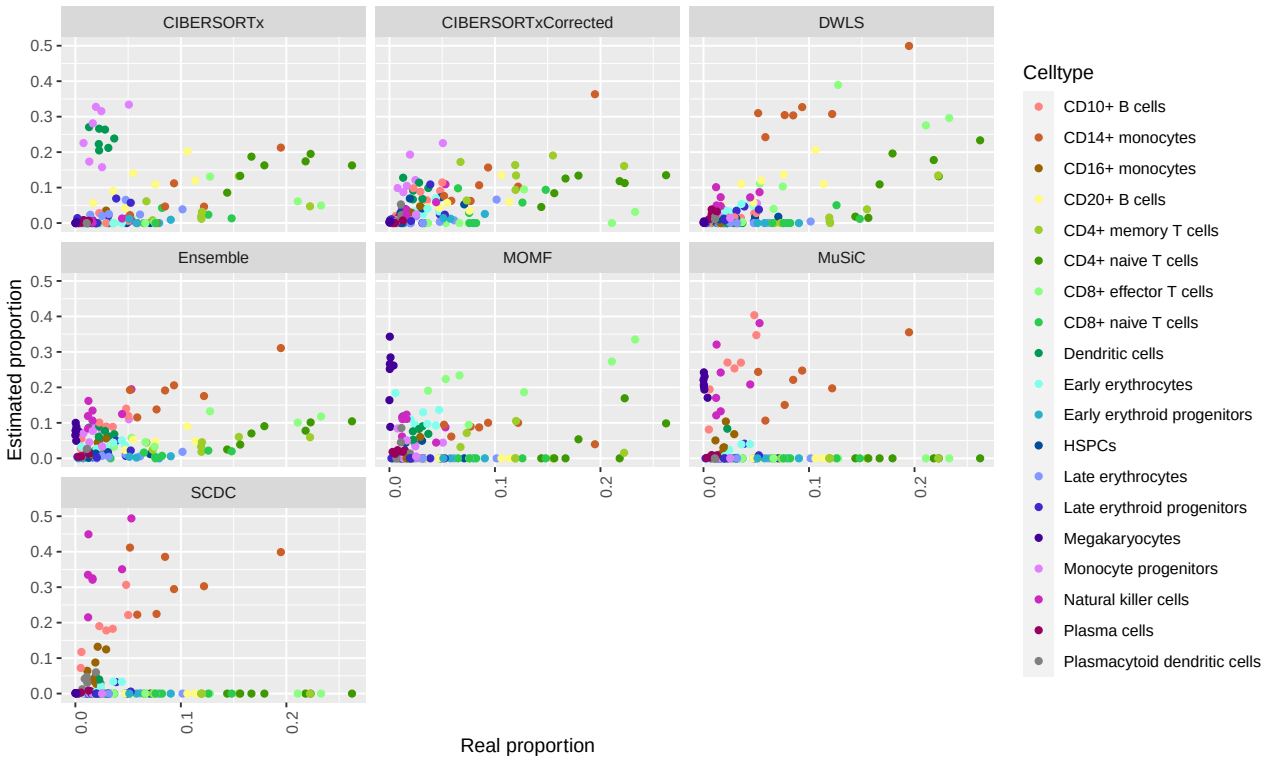
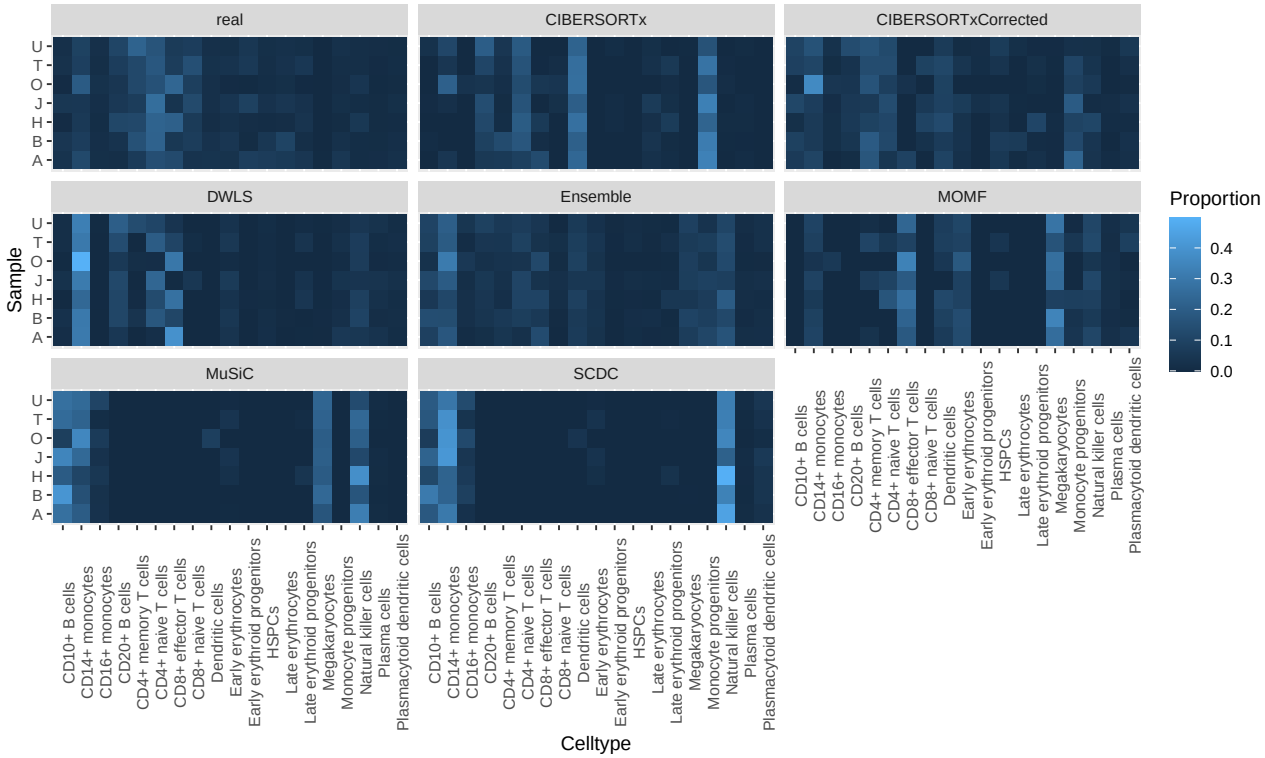


Figure 6: Validation of similarity between bulk and scRNA data. Log scaled bulkRNA-Seq expression values are shown on the x-axis while log scaled scRNA-Seq expression values are shown on the x-axis.



(a) Scatter plot



(b) Heatmap

Figure 7: Benchmark results on BM data. The "real" proportions are based on the scRNA data which was shown to be close to the flow cytometry proportions.

Table 3: The computational performance of the methods. The CPU time is the time the CPU spent on the deconvolution and the memory is the peak memory usage of the deconvolution.

Method	CPU time (HH:MM:SS)	Memory
CIBERSORTx	00:09:27	15.5 GB
CIBERSORTxCorrected	00:25:58	29.8 GB
DWLS	07:22:41	84.0 GB
MOMF	53:38:19	<22.0 GB
MuSiC	00:04:34	<21.9 GB
SCDC	00:13:52	<22.8 GB

was the fastest method followed by CIBERSORTx without batch correction and SCDC. Both DWLS and especially MOMF took much longer. In terms of memory, most methods used around 20 GB of memory. DWLS however, needed 84 GB. Note that MOMF, MuSiC and SCDC, the peak memory useage was achieved while preparing the input data before running the methods themselves. Their memory usage would be lower if the input data was read in a single step.

One of the biggest outliers in the deconvolution of the bone marrow data was the high estimated proportion for the megakaryocytes. It was hypothesised that the cause of this is the high error in the estimation of the mean expression for this cell type as a result of the lower number of cells (43) for this cell type. This could have caused the mean expression to be too close to another cell type. However, the correlation of megakaryocytes with other cell types was relatively low in the scRNA dataset used in the deconvolution (figure 8). The maximum correlation was 0.8 with natural killer cells while most cell types had a correlation of more than 0.9 with at least one other cell type. Furthermore, even for two cell types with a correlation of 0.99, repeatedly sampling 50 cells results in two distinct groups of means (figure 9).

The large overestimation of the proportion of megakaryocytes shows that similar inaccuracies are also possible for the PKD data. Additionally, another common inaccuracy in the deconvolution is that many cell types have an estimated proportions of zero or close to zero despite having a higher real proportion. This also means that cell types with a predicted proportios of zero in the PKD data are likely to have a higher proportion in reality.

3.3 PKD

CIBERSORTx, MuSiC and DWLS were applied on the PKD bulk data and kidney scRNA data. These 3 methods were chosen because they were the most accurate for the simulated and the bone marrow data. The results of the deconvolution on the PKD data is shown in figure 10. CIBERSORTx with batch correction was also applied on the data but the results were not realistic mainly because the proportion of T lymphocytes was estimated at 30% (appendix C).

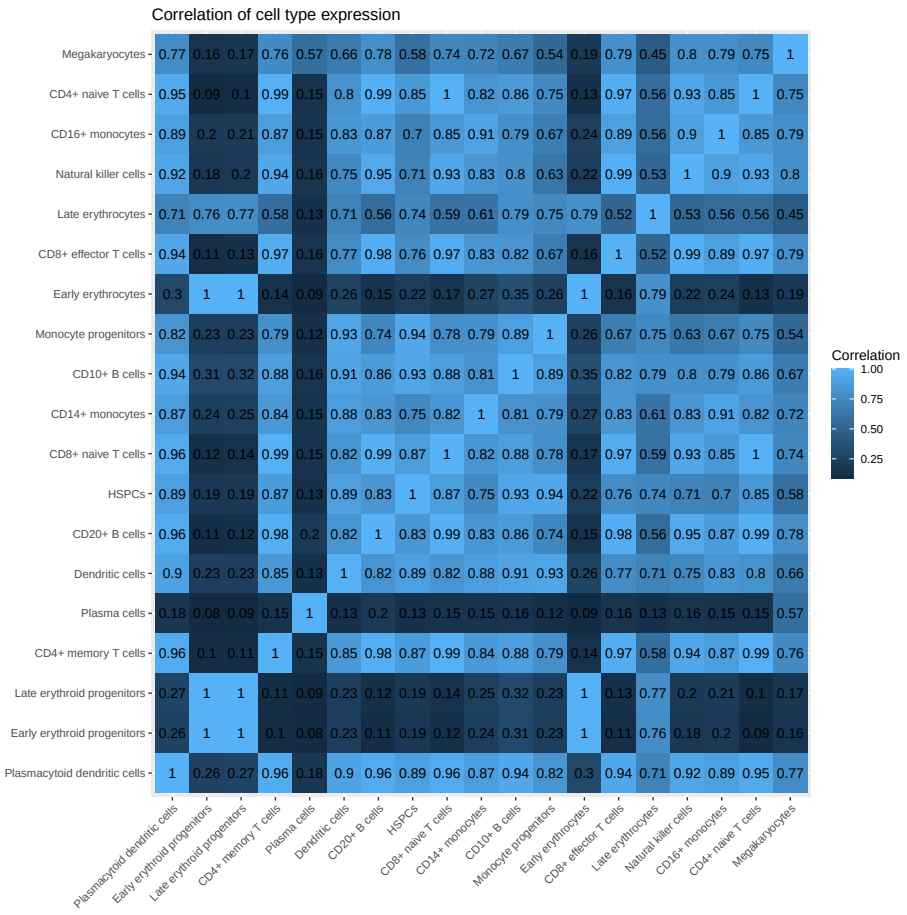


Figure 8: The Pearson correlation for every pair of cell types in the bone marrow scRNA data.

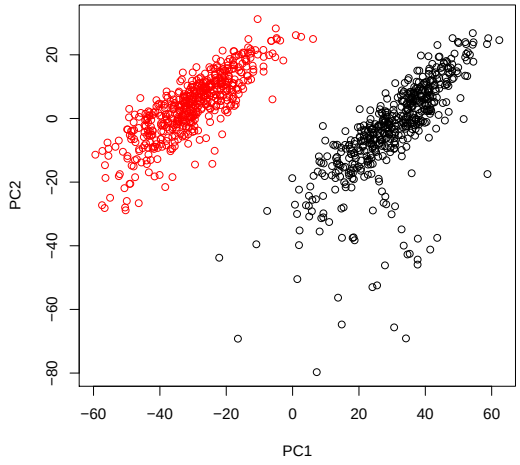


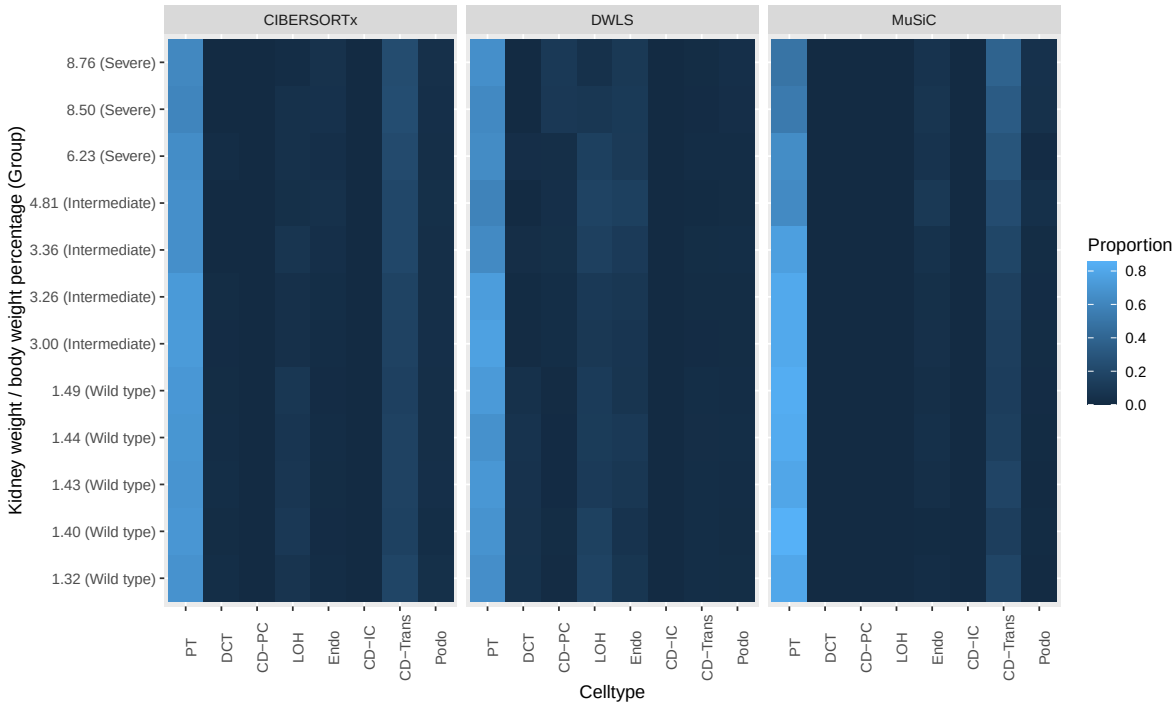
Figure 9: Mean expression CD4+ naive T cells and CD20+ B cells. Each point is the mean expression of the cell type based on 50 sampled cells out of 1000. The dimensionality was reduced with PCA.

For the three methods, Proximal tubules were the cell type with the highest proportion. Additionally, the proportion decreases as the disease severity is worse. For the immune cells, only macrophages and fibroblasts are clearly visible. MuSiC only detected macrophages while CIBERSORTx and DWLS also detected fibroblasts. When detected, both macrophages and fibroblasts had increased proportions in mice with a more severe progression of PKD.

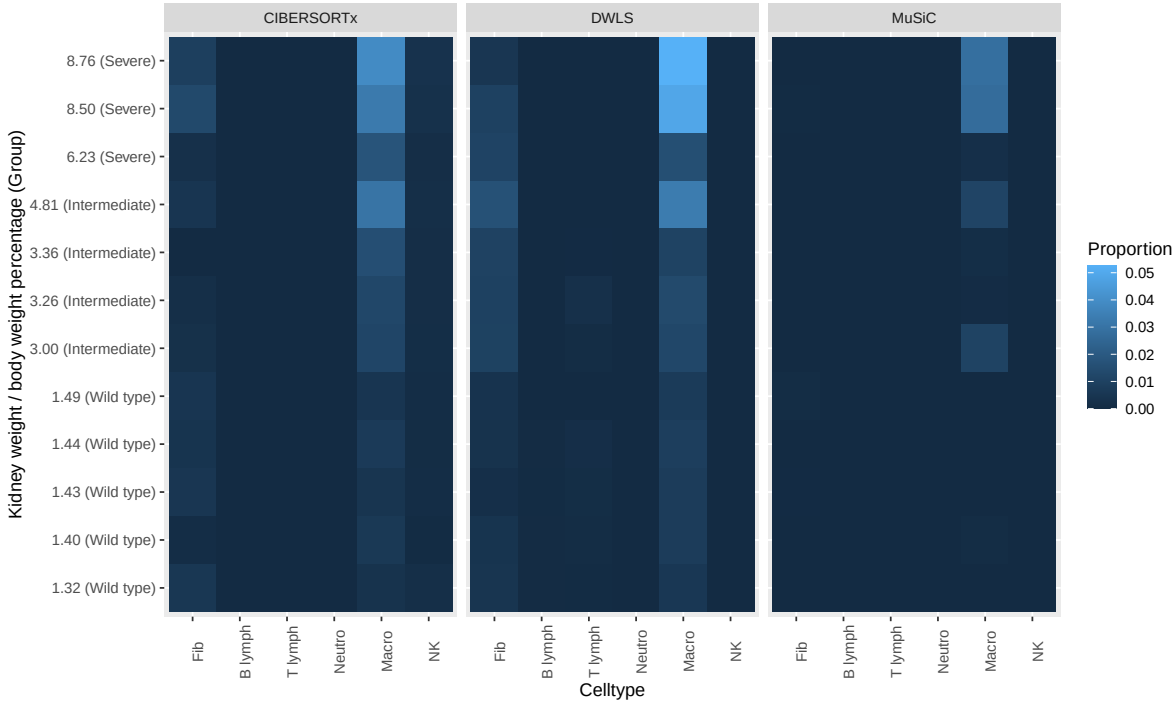
3.4 Convolution

In order to estimate the changes in bulk gene expression as a result of the changes in cell type proportion, cell type specific gene expression was multiplied with the proportions estimated by DWLS in wild-type mice and also in mice with a severe PKD phenotype. In order to attribute these gene expression changes to a single cell type, this was repeated for vectors of cell type proportions where only one cell type is changed from the wild-type to severe PKD proportion.

Two small sections of the convolution results are shown in figure 11. In the first part, some marker genes of macrophages (C1qa, C1qb and C1qc) are estimated to have a five-fold increase in expression in the PKD samples as a result of the changes in cell type proportions. This fold change can be explained by only the increase in macrophages. Similarly, a marker gene for fibroblasts (S100a4) has an almost two-fold increase in gene expression. In this case, most of the change can be explained by the increase proportions of fibroblasts.



(a) Epithelial and endothelial cells



(b) Immune cells and fibroblasts

Figure 10: Predicted proportions in PKD. The kidney weight to body weight percentage and phenotype label are included in the labels on the y-axis.

Cell type proportion change	All -	PT Decrease	DCT Decrease	CD-PC Increase	Fib Increase	B lymph Decrease	T lymph Decrease	LOH Decrease	Endo Increase	CD-IC Decrease	Neutro Decrease	Macro Increase	NK Decrease	CD-Trans Decrease	Podo Increase
Luzp1	1,082	1,071	0,976	1,056	0,998	1,000	1,001	0,979	1,061	1,000	1,000	0,992	1,000	0,984	1,014
Kdm1a	1,042	1,045	0,987	1,084	1,002	1,000	1,000	0,992	1,022	1,000	1,000	0,985	1,000	0,958	0,997
Ephb2	0,951	1,144	1,030	1,173	0,996	1,000	1,001	0,755	0,984	1,000	1,000	0,968	1,000	0,974	0,992
C1qb	5,239	1,143	1,044	0,932	0,999	1,000	1,001	1,045	0,975	1,000	1,000	5,235	1,000	1,007	0,992
C1qc	5,293	1,143	1,044	0,933	0,996	1,000	1,001	1,046	0,973	1,000	1,000	5,291	1,000	1,008	0,992
C1qa	5,305	1,144	1,044	0,932	0,996	1,000	1,001	1,045	0,973	1,000	1,000	5,305	1,000	1,008	0,992
Zbtb40	1,004	0,985	1,017	0,973	0,999	1,000	1,000	1,017	1,000	1,000	1,000	1,011	1,000	0,999	0,997
Wnt4	1,475	1,122	1,019	1,435	1,002	1,000	1,000	1,033	1,038	1,000	1,000	1,013	1,000	0,849	1,057
Cdc42	1,119	1,022	1,007	1,038	1,017	1,000	0,999	1,012	1,020	1,000	1,000	1,035	1,000	0,981	1,008
S100a16	1,187	1,060	1,011	1,090	0,996	1,000	1,001	1,013	1,103	1,000	1,000	0,974	1,000	0,988	1,004
S100a3	1,410	1,135	1,014	1,084	0,995	1,000	0,997	1,047	1,291	1,000	1,000	0,968	1,000	1,008	0,992
S100a4	1,842	1,139	1,043	0,932	1,721	1,000	0,981	1,043	1,124	1,000	1,000	0,975	0,999	1,008	0,992
S100a6	1,372	1,144	1,044	0,934	1,061	1,000	0,998	1,046	1,303	1,000	1,000	0,971	1,000	1,007	0,994
S100a7a	0,772	1,114	0,957	0,932	0,995	1,000	1,000	0,835	1,026	1,000	1,000	0,968	1,000	1,008	0,992

Figure 11: Two small parts of the resulting matrix of the convolution. The numbers are expected fold changes in severe PKD samples versus wild type samples as a result of changes in cell type proportions. The first column represents the fold change if all proportions change. The rest of the columns represent the fold change if only the proportion of that cell type changes.

4 Discussion

Six method for expression deconvolution were benchmarked. MuSiC, CIBERSORTx and DWLS performed the best on the simulated data while CIBERSORTx (with batch correction) and DWLS performed the best on the bone marrow data. CIBERSORTx with batch correction, however, had unexpected results for the PKD data. One possible reason for this is that the batch correction in CIBERSORTx is based on the results of deconvolution without batch correction. Inaccuracies in this deconvolution iteration could result in an inaccurate batch correction which in turn leads to an inaccurate second deconvolution iteration.

In terms of computational peformance, MuSiC, SCDC and CIBERSORTx were relatively fast with low memory usage while MOMF and DWLS required much more CPU time and memory. For this reason, CIBERSORTx seems to be the best method for situations where the computational resources are limited. In situations where this is not an issue, DWLS seems to offer a slight advantage over CIBERSORTx in terms of accuracy.

For some of the methods, there are some limitations in the benchmarks. For example, SCDC was used with only a single scRNA reference dataset and could therefore not take advantage of the ensemble model. This probably did not impact the performance, because the used reference should already the ideal reference since it is generated from the same samples. MuSiC was used without the tree-guided approach because this required additional information outside of the scRNA dataset. Finally, the trainings set data for DigitalDLSorter that is based on prior knowledge was replace with trainings data not based on prior knowledge. DigitalDLSorter should therefore perform better in situations where prior knowledge can be used because the training data will be more representative of the real data.

When applied on the PKD dataset, the best methods were able to detect some disease relevant changes like the increase in fibroblasts, macrophages and endothelial cells. For example, inflammation is involved in PKD. Cultured PKD1 deficient cells express Mcp1 and Cxcl16 which attract macrophages that

in turn promote cyst growth [5]. This is also directly related to the altered IL-1 signaling found in PKD using RNA-Seq [9]. Similarly, basic fibroblast growth factor (bFGF or FGF2) [21, 8] and vascular endothelial growth factor (VEGF) [8] are upregulated in PKD. These growth factors promote the growth of fibroblasts and vascular endothelial cells. Like MuSiC, CIBERSORTx and DWLS also estimated a decrease in proximal tubular cells. However, this decrease was far less strong for CIBERSORTx and DWLS than for MuSiC. This likely means that the decrease in proximal tubular cells is at least partially an artefact of MuSiC caused by dedifferentiation of the proximal tubular cells. For CIBERSORTx and DWLS, about half of the remaining change in proportion can be explained by the increase in macrophages and fibroblasts. The three methods failed to reliable detect some of the cell types including the T lymphocytes, B lymphocytes and natural killer cells. Additionally, for the collecting duct cells, the methods were not able to reliably differentiate between principal cells, intercalated cells and the cells that are transitioning between the two cell types.

The changes in gene expression of some genes can be attributed to changes in cell type proportions with the convolution method. The accuracy, however, is dependent on both the accuracy of the estimated proportions and also on the similarity of the gene expression of the bulk and the scRNA. For example, if a cell type marker exists in the bulkRNA data but not in the scRNA data, the fold change cannot be correctly estimated.

In order to estimate the cell type specific gene expression in bulkRNA samples, it is possible to fix the matrix with the bulk expression and the vector with the cell type proportions in the equation for expression deconvolution. CIBERSORTx is one method that has this ability. This, however, requires a large number of samples in order to get realistic results. Newman et al. [15] state that the largest gains in accuracy were achieved when analyzing at least four-fivefold more mixture samples than cell types. Analyzing a large group of PKD mixture samples in order to gain a deeper understanding of gene expression changes in individual cell types would be very interesting.

Future research could also be done on improving deconvolution methods by investigating the factors that cause deconvolution to give inaccurate results. For example, having more cell types or having more similar cell types should make the deconvolution less accurate while having more cells per cell type or more reads in the RNA-Seq data should make the deconvolution more accurate. It could also be interesting to investigate if linear combinations of certain cell types can be confused with linear combinations of other cell types.

5 Conclusion

MuSiC, SCDC, MOMF, CIBERSORTx, digitalDLsorter and DWLS were benchmarked for both a simulated and a experimental bone marrow dataset. DWLS and CIBERSORTx performed the best in the benchmarks. DWLS offered good performance at a high computational cost and CIBERSORTx offered almost as good performance at a far lower computational cost. DWLS was also able to detect more rare cell types because of the weighting scheme that is focused on rare cell types. On datasets with a low complexity (e.g. low number of cell types), MuSiC also performed very well and is faster than the other methods. The benchmark on the bone marrow data showed that although deconvolution on an experimental complex dataset recovers some of the real changes in proportions, it is limited in accuracy.

MuSiC, CIBERSORTx and DWLS were applied on a PKD dataset in order to find changes in cell proportions associated with the disease. Similarly to the earlier results of MuSiC, an increased proportion of macrophages for mice with a more severe disease phenotype was detected. Unlike MuSiC, CIBERSORTx and DWLS also estimated an increased proportion of fibroblasts. Both these results seem to confirm the involvement of macrophages and fibroblasts in PKD. Another cell proportion change detected by MuSiC was the decrease in proportion of proximal tubular cells. This change was much smaller in CIBERSORTx and DWLS and indicates that the decrease of proximal tubular cells is at least partially an artefact of the deconvolution method.

Finally, the proportions estimated by DWLS were used to attribute the changes in gene expression of some cells to the changes in cell type proportions. Macrophage and fibroblast marker genes were estimated to have an increase in gene expression solely because of an increase in proportion of macrophages and fibroblasts. It can be useful to combine these results with differential expression testing in PKD, because it can show when the differential expression is likely a result of the changes in cell type proportions.

References

[1] Carsten Bergmann et al. "Polycystic kidney disease". In: *Nature Reviews Disease Primers* (2018). DOI: 10.1038/s41572-018-0047-y.

[2] Seokho Kim et al. "The polycystin complex mediates Wnt/Ca2+ signalling". In: *Nature Cell Biology* 18.7 (2016), pp. 752–764. ISSN: 14764679. DOI: 10.1038/ncb3363.

[3] Kevin Retailleau and Fabrice Duprat. "Polycystins and partners: Proposed role in mechanosensitivity". In: *Journal of Physiology* 592.12 (2014), pp. 2453–2471. ISSN: 14697793. DOI: 10.1113/jphysiol.2014.271346.

[4] Chiara Formica and Dorien J.M. Peters. "Molecular pathways involved in injury-repair and ADPKD progression". In: *Cellular Signalling* 72.January (2020), p. 109648. ISSN: 18733913. DOI: 10.1016/j.cellsig.2020.109648. URL: https://doi.org/10.1016/j.cellsig.2020.109648.

[5] Anil Karihaloo et al. "Macrophages promote cyst growth in polycystic kidney disease". In: *Journal of the American Society of Nephrology* 22.10 (2011), pp. 1809–1814. ISSN: 10466673. DOI: 10.1681/ASN.2011010084.

[6] Kurt A. Zimmerman et al. "Tissue-resident macrophages promote renal cystic disease". In: *Journal of the American Society of Nephrology* 30.10 (2019), pp. 1841–1856. ISSN: 15333450. DOI: 10.1681/ASN.2018080810.

[7] Michel Chonchol et al. "Fibroblast growth factor 23 and kidney disease progression in autosomal dominant polycystic kidney disease". In: *Clinical Journal of the American Society of Nephrology* 12.9 (2017), pp. 1461–1469. ISSN: 1555905X. DOI: 10.2215/CJN.12821216.

[8] Xuewen Song et al. "Systems biology of autosomal dominant polycystic kidney disease (ADPKD): Computational identification of gene expression pathways and integrated regulatory networks". In: *Human Molecular Genetics* 18.13 (2009), pp. 2328–2343. ISSN: 09646906. DOI: 10.1093/hmg/ddp165.

[9] Tareq B. Malas et al. "Prioritization of novel ADPKD drug candidates from disease-stage specific gene expression profiles". In: *EBioMedicine* 51 (2020). ISSN: 23523964. DOI: 10.1016/j.ebiom.2019.11.046.

[10] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental and Molecular Medicine* 50.8 (2018). ISSN: 20926413. DOI: 10.1038/s12276-018-0071-8. URL: http://dx.doi.org/10.1038/s12276-018-0071-8.

[11] Haojia Wu et al. "Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: Rare cell types and novel cell states revealed in fibrosis". In: *Journal of the American Society of Nephrology* 30.1 (2019), pp. 23–32. ISSN: 15333450. DOI: 10.1681/ASN.2018090912.

[12] Xuran Wang et al. "Bulk tissue cell type deconvolution with multi-subject single-cell expression reference". In: *Nature Communications* 10.1 (Dec. 2019). ISSN: 20411723. DOI: 10.1038/s41467-018-08023-x.

[13] Meichen Dong et al. "SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references". In: *Briefings in Bioinformatics* (Jan. 2020). ISSN: 1477-4054. DOI: 10.1093/bib/bbz166.

[14] Xifang Sun, Shiquan Sun, and Sheng Yang. "An Efficient and Flexible Method for Deconvoluting Bulk RNA-Seq Data with Single-Cell RNA-Seq Data". In: *Cells* 8.10 (Sept. 2019), p. 1161. ISSN: 2073-4409. DOI: 10.3390/cells8101161.

[15] Aaron M. Newman et al. "Determining cell type abundance and expression from bulk tissues with digital cytometry". In: *Nature Biotechnology* (July 2019). ISSN: 15461696. DOI: 10.1038/s41587-019-0114-2.

[16] Carlos Torroja and Fatima Sanchez-Cabo. "Digitaldlsorter: Deep-learning on scrna-seq to deconvolute gene expression data". In: *Frontiers in Genetics* 10.OCT (2019). ISSN: 16648021. DOI: 10.3389/fgene.2019.00978.

[17] Daphne Tsoucas et al. "Accurate estimation of cell-type composition from gene expression data". In: *Nature Communications* 10.1 (2019). ISSN: 20411723. DOI: 10.1038/s41467-019-10802-z. URL: http://dx.doi.org/10.1038/s41467-019-10802-z.

[18] Karolyn A. Oetjen et al. "Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry". In: *JCI insight* 3.23 (2018). ISSN: 23793708. DOI: 10.1172/jci.insight.124928.

[19] Jihwan Park et al. "Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease". In: *Science* 360.6390 (2018), pp. 758–763. ISSN: 10959203. DOI: 10.1126/science.aar2131.

[20] Andrew D. Yates et al. "Ensembl 2020". In: *Nucleic Acids Research* 48.D1 (2020), pp. D682–D688. ISSN: 13624962. DOI: 10.1093/nar/gkz966.

[21] T Nakamura et al. "Growth Factor Gene Expression in Kidney of Murine Polycystic Kidney Disease". In: *Journal of the American Society of Nephrology* 3.7 (1993), pp. 1378 –1386.

A Convolution example

Cell type gene expression

	Macro	PT	CD
Gene 1	2	0	0
Gene 2	0	2	0
Gene 3	0	0	2
Gene 4	4	4	4
Gene 5	4	4	4

Estimated proportions

	wild type	Disease	Macro-changed	PT-changed	CD-changed
Macro	0,05	0,10	0,10	0,05	0,05
PT	0,48	0,45	0,45	0,45	0,50
CD	0,48	0,45	0,45	0,50	0,45
SUM	1	1	1	1	1

Simulated bulk expression

wild type

Gene 1	0,1
Gene 2	0,95
Gene 3	0,95
Gene 4	4
Gene 5	4

Disease (All-changed)

Gene 1	0,2
Gene 2	0,9
Gene 3	0,9
Gene 4	4
Gene 5	4

Macro-changed

Gene 1	0,2
Gene 2	0,9
Gene 3	0,9
Gene 4	4
Gene 5	4

PT-changed

Gene 1	0,105
Gene 2	0,9
Gene 3	0,995
Gene 4	4
Gene 5	4

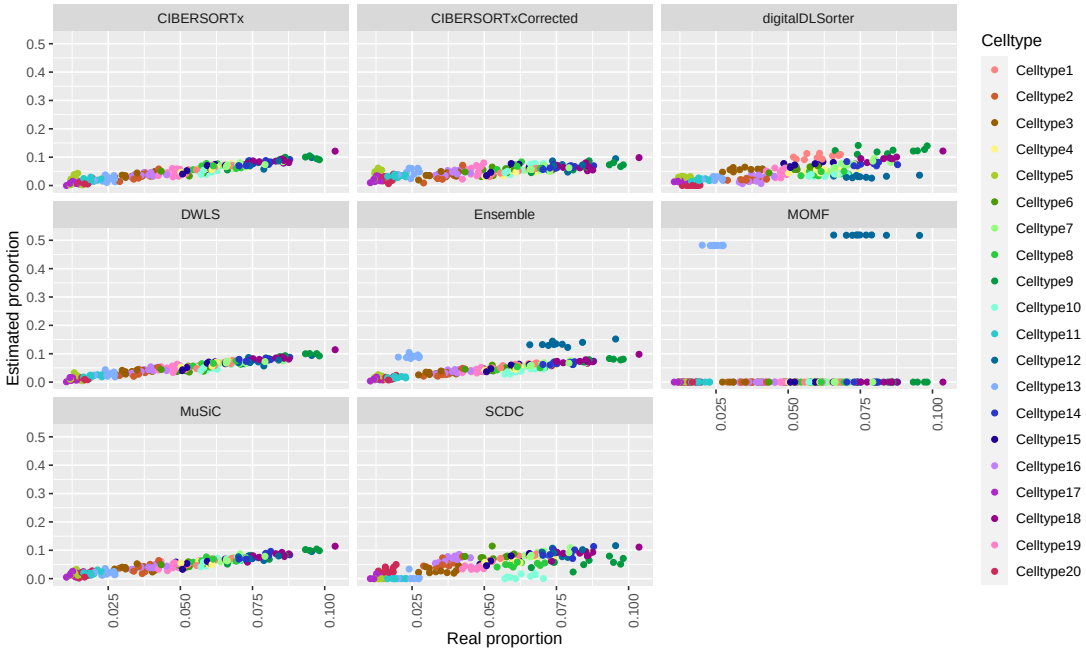
CD-changed

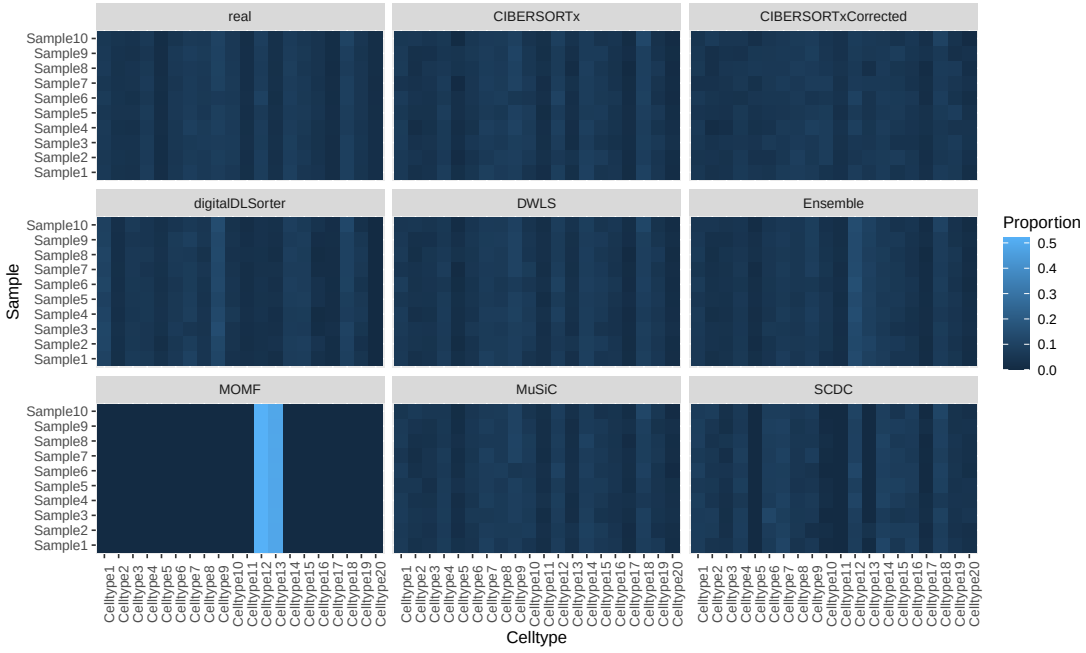
Gene 1	0,1
Gene 2	1
Gene 3	0,9
Gene 4	4
Gene 5	4

Fold changes (bulk with x cell type proportion changed / bulk wild type)

	All	Macro	PT	CD
Cell type proportion change	-	Increase	Decrease	Decrease
Gene 1	2,000	2,000	1,048	1,048
Gene 2	0,947	0,947	0,947	1,048
Gene 3	0,947	0,947	1,048	0,947
Gene 4	1	1	1	1
Gene 5	1	1	1	1

B Full simulation results





C Full PKD results