Evaluation of computational pipelines for multimodal single-cell data

Jonathan Haab
johaab@student.ethz.ch

Supervisor
A. Sonrel
Robinson lab UZH

Mentor
Prof. Dr. D. Iber
CoBi ETHZ

Group leader
Prof. Dr. M. D. Robinson
Robinson lab UZH

A report presented as part of the
Master in Computational Biology and Bioinformatics
(CBB)

D-BSSE
ETHZ
Switzerland
Spring 2021

Abstract

Less than ten years after the first scRNA-seq study was published, the field of single-cell analysis took an other big step forward with the advent of simultaneous measurement of gene and protein expression levels. Even though methods like CITE-seq and REAP-seq represent a major technical improvement, it is not clear how the full potential of such data will be unravelled. In this project, we aimed at evaluating the performance of different normalization methods and their combination with alternative clustering algorithm on CITE-seq data. We investigated methods specifically developed for CITE-seq data processing already available and we built our own metric to asses the performance of normalization applied on ADT counts in the absence of ground truth. We found that ADTs information alone leads most of the time to poor clustering and that more effort should be invested to combine the proteomic and transcriptomic signals efficiently.

1 Introduction

1.1 Single-cell sequencing

The cell is the fundamental unit in Biology and the resolution attained for its observation has not ceased to increase during the past decades. One notable breakthrough was the development of single-cell sequencing techniques about which the very first study was published in 2009 [6]. Single-cell sequencing can reveal cell-to-cell variability in populations which might have been considered as homogeneous beforehand and is important to understand the role of individual cells in their system-level function. It has been used extensively to study intra-tumor genetic heterogeneity in cancer development or treatment response [19].

1.2 scRNA-seq

Sequencing RNA is of particular interest when studying the variation in gene expression of a cell population rather than simply its genetic content. The variation in gene expression is linked to a certain extent with cell function and phenotype, even though the exact relationship remains unclear in some cases. This technique focuses on messenger RNA (mRNA) which correlates well with functional proteins to infer the cell phenotype from RNA counts. scRNA-seq represents a tremendous advancement compared to microarrays largely used in the past as it is usually more reliable for measuring gene expression level changes [13]. The general workflow of scRNA-seq is as follow: First, single cells are isolated from a tissue sample and then lysed so that their mRNA content becomes accessible. Poly[T] sequence primers are used thereafter to capture mRNA molecules by binding to their poly[A] tails and then reverse transcriptase (RT) converts them into complementary DNA (cDNA). Later on, cDNA is amplified, usually using polymerase chain reaction (PCR), indexed and pooled for sequencing. Once the sequencing is done, bioinformatics is applied to perform quality control (QC), discriminate biological variation from technical noise and, finally, interpret the data [9].

1.3 CITE-seq

Like single-cell RNA sequencing (scRNA-seq) before it, cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) increased tremendously the resolution with which cell populations can be observed, raising new interrogations and encouraging exciting discoveries. The strength of this modern method first described in 2017 [11] holds in the expansion of the information gathered to include phenotypic profiling of the cells. CITE-seq is a method which combines detection of mRNA and protein, boosting scRNA-seq capability. In addition to the scRNA-seq workflow described above, this technique relies on antibody-derived tags (ADTs) [11]. Those tags were first constructed by conjugating some of the monoclonal antibodies typically used in flow cytometry to DNA oligonucleotides and a poly[A] tail. The oligonucleotides are used as barcodes, i.e. a unique barcode is associated with a specific protein, and protein count is determined via the number of corresponding tags detected. The poly[A] tail is used to convert and ultimately amplify the barcode along with the mRNA, so that the same workflow is able to simultaneously gather information about the RNA and protein content of a cell. An other method, called RNA expression and protein sequencing (REAP-seq), provides essentially the same information as CITE-seq but differs in the way the DNA barcode is conjugated to the antibody [10].

This method opened exciting perspectives as it integrates information about the actual protein content into the well established scRNA-seq methodology. The authors who presented it claimed that it represents an alternative to flow cytometry, considered as the "gold standard" to identify cell types [11], even though the later technique remains cheaper, faster and more precise at the moment. One interesting advantage of CITE-seq over fluorescence-activated cell sorting (FACS) is that the former is not limited by spectral interference [18].

1.4 Our contribution

In this report, we describe a first attempt to evaluate the performance of methods initially developed for scRNA-seq data and their combination on different datasets as well as investigate novel methods specifically designed for CITE-seq data. As can be seen in Table 1, different methods are employed to analyse CITE-seq data and no clear consensus exists as of today. Therefore, we used pipeComp, a flexible framework for pipeline comparison, to shed light on what seems to be the best approach for CITE-seq data analysis. The structure of the pipelines used was pretty

standard: the data were filtered, normalized and then clustered. A single filtering method was used but several alternative normalization were compared in combination with two different clustering algorithms. In addition to this benchmark, we sought a novel way to evaluate the performance of normalization methods in absence of ground truth by taking advantage of isotypes and positive controls present among the protein counts.

In the end, we highlighted the limitations of ADTs to group the cell with respect to their cell type. We depreciated the use of CLR_{ADTs} normalization and pointed out problems that may arise when relying on an arguably strict ground truth. Moreover, we proposed ways to build on and extend the work presented in this project.

2 Method

2.1 Data

The data from five different studies were used and are summarized in Table 1. They were essentially counts matrices of both mRNA and ADT for all sampled cells. Most data described peripheral blood mononuclear cells (PBMCs) since blood is a very convenient tissue to collect from donors. Krebs et al. also provided samples from Kidney CD3+ T cells. Some cells and genes had to be discarded prior to our analysis because no ground truth was available or because the conversion of gene symbols to ensembl IDs failed, which were required for the different datasets to be compatible with our pipeline. Note that the large number of cells found in Hao et al. caused the pipelines run to exceed time and memory resources at hand. More specifically, the clustering step was found to be problematic. This dataset was discarded from the main analysis.

2.2 Filtering

The same filtering technique was used for every alternative pipeline, following the workflow presented by Amezquita et al. [15]. More precisely, doublets were detected and discarded using the scDblFinder function from Germain et al. [20], quality control metrics were computed for each cell with additional focus on mitochondrial genes and low quality cells were discarded. Additional outlier cells were discarded based on the median-absolute-deviation (MAD) of their ADT counts in the end.

2.3 Normalization

The ADT counts were normalized using different approaches, allowing for comparison later on. In the following, $\mathbf{X}$ represents the ADT counts matrix, n the number of ADTs and m the number of cells.

logNormCounts (log-transformation) is the first normalization we used. This method is commonly used for scRNA-seq and is designed to transform data initially log-normal distributed into a Gaussian distribution, which is useful for later clustering or dimensionality reduction possibly relying on the Euclidean distance. Prior to the transformation, the count of each ADT was divided by the size factor for a particular cell so that the whole process returns log-transformed normalized expression values in the end. In mathematical form:

$$\text{LT}(X_{i,j}) = \log\left[\frac{X_{i,j}}{\sum X_{\cdot,j}} + 1\right]$$
$$\forall\ 0 \leq i < n,\ 0 \leq j < m$$

where $\sum X_{\cdot,j}$ is the sum over every row of the jth column.

CLR (centered log-ratio transformation) uses the geometric mean across features (alt. 1), i.e. the ADTs in our case, or across cells (alt. 2) as reference for its transformation. One advantage of this widely used method is that the one-to-one difference between features value is preserved even if some counts are discarded [7].

$$\text{alt. 1} \qquad \text{CLR}_{\text{ADTs}}(X_{i,j}) = \log\left[\frac{X_{i,j}}{g(X_{\cdot,j})}\right]$$
$$\forall\ 0 \leq i < n,\ 0 \leq j < m$$

where $g(X_{\cdot,j}) = (\prod X_{\cdot,j})^{\frac{1}{n}}$

$$\text{alt. 2} \qquad \text{CLR}_{\text{Cells}}(X_{i,j}) = \log\left[\frac{X_{i,j}}{g(X_{i,\cdot})}\right]$$
$$\forall\ 0 \leq i < n,\ 0 \leq j < m$$

where $g(X_{i,\cdot}) = (\prod X_{i,\cdot})^{\frac{1}{m}}$

A variation of CLR across cells was also applied using the ScaleData function from Satija et al. [8]. That method extends CLR by scaling and centering the $\text{CLR}_{\text{cells}}$ transformed data. The centering is achieved by subtracting the average expression for every features and the scaling by dividing the centered feature expression levels with their standard deviations.

$$\text{scaleData}(X_{i,j}) = \frac{X'_{i,j} - \bar{X}'_i}{\sigma_i}$$
$$\forall\ 0 \leq i < n,\ 0 \leq j < m$$

where $X'_{i,j} = \text{CLR}_{\text{Cells}}(X_{i,j})$, $\bar{X}'_i$ is the average and σ_i the standard deviation of the expression of the ith ADT.

DSB (Denoised and scaled by background) was specially developed to normalized CITE-seq data and relies on empty droplets. Those droplets allow to assess the substantial amount of background signal generated by unbound ADTs [18]. Unfiltered data must be available in order to use DSB, as the droplets that would usually be discarded during data processing contain crucial information to assess background noise. For each cell, the DSB normalized counts are computed by taking the logarithm of the count, subtracting the mean expression of empty droplets and dividing the whole by the standard deviation of the signal from empty droplets.

Authors	Cell type	Number of cells	Number of genes	Number of ADTs	Normalization
Granja et al.	PBMCs	14,804	20,074	21	CLR
Hao et al.	WBCs	94,674	20,421	228	CLR
Kotliarov et al.	PBMCs	53,201	20,759	87	DSB
Krebs et al.	PBMCs CD3$^+$ T cells	6,377	12,312	17	CLR
Krebs et al.	Kidney CD3$^+$ T cells	3,813	12,538	17	CLR
Mair et al.	PBMCs	27,242	488	42	LT

Table 1: The table above summarizes the datasets used in our analysis. PBMCs stands for peripheral blood mononuclear cells and WBCs for white blood cells. The number of cells, genes and ADTs presented here might be reduced compared to the data the authors used in their study for compatible reason with our pipelines. It was essential to gather datasets containing different range of cells and ADTs and different proportion of genes and ADTs. The normalization used by the authors was reported as it influenced their clustering which we accepted as ground truth.

$$\mathrm{DSB}(X_{i,j}) = \frac{\log(X_{i,j}+10) - \mu_e}{\sigma_e}$$
$$\forall\ 0 \le i < n,\ 0 \le j < m$$

where μ_e is the mean expression of empty droplets and σ_e their standard deviation.

A second, optional step can be performed where each cell is denoised to account for technical component (i.e. per-droplet differences in oligo tag capture efficiency and cell-specific differences in non-specific antibody staining). A Gaussian mixture model is fit on the transformed counts of each cell with two mixture components to do so.

2.4 PICS score

Most scoring methods like silhouette coefficient, mutual information (MI) or rand index (RI) rely on ground truth for their computation. As an effort to move away from this dependency and taking advantage of isotype and positive controls present among the ADTs used, we developed and tested our positive and isotype controls separation (PICS) score. This novel approach uses both the isotype and positive control distributions after normalization, without requiring any sort of ground truth. This score compares how well the normalization manages to bring down signal from isotypes while preserving signal from positive controls. Initially, the distributions of normalized isotype and positive control ADT counts are shifted so that the minimum normalized count is set on zero and scaled so that the maximum is set on one. This step ensures the standardization of normalized counts distribution, as normalization methods may return values of different scales. Then the score is computed in two stages. First, the isotype distribution is evaluated alone to assess its compactness on zero. This compactness is assessed using the median of the isotype distribution median($\mathbf{I}$) and the standard deviation of the same distribution $s_{\mathbf{I}}$. The inverse of the product of median($\mathbf{I}$) and $s_{\mathbf{I}}$ is taken since a median close to zero and a small standard deviation should lead to a high score. A small constant added to the denominator prevents undefined solution. The value of that constant was chosen in order for the two parts of the PICS score to be around the same scale, hence preventing one part to dominate the total score. We chose a value of 0.1, leading to a maximum $\mathrm{PICS}_{\mathrm{neg}}$ of 10 in case either median($\mathbf{I}$) or $s_{\mathbf{I}}$ happened to be zero. Second, the separation between the isotype and the positive control distribution is computed by a t-test returning the t statistic. The t statistic depends on the mean of both the positive and isotype distribution as well as their variance and the number of cells observed. The larger the difference between the two distribution, the higher the score. In the end, both subscores are multiplied together.

$$\mathrm{PICS}(\mathbf{I},\mathbf{P}) = \mathrm{PICS}_{\mathrm{neg}}(\mathbf{I}) \cdot \mathrm{PICS}_{\mathrm{pos}}(\mathbf{I},\mathbf{P})$$
$$\mathrm{PICS}_{\mathrm{neg}}(\mathbf{I}) = \frac{1}{\mathrm{median}(\mathbf{I}) \cdot s_{\mathbf{I}} + 0.1}$$
$$\mathrm{PICS}_{\mathrm{pos}}(\mathbf{I},\mathbf{P}) = t_{\mathbf{I},\mathbf{P}} = \frac{\bar{\mathbf{I}} - \bar{\mathbf{P}}}{\sqrt{\frac{s_{\mathbf{I}}^2}{n_{\mathbf{I}}} + \frac{s_{\mathbf{P}}^2}{n_{\mathbf{P}}}}}$$

where $s_{\mathbf{I}}$ and $s_{\mathbf{P}}$ are the estimated standard deviation of the isotype and positive control distribution, respectively. $n_{\mathbf{I}}$ and $n_{\mathbf{P}}$ are the number of cells, which are identical in our case so that $n_{\mathbf{I}} = n_{\mathbf{P}}$. Note that because of the absence of explicit positive control ADTs, anti-CD45 were selected as this marker is expressed on all leukocytes [1]. If multiple positive and negative controls are found, the final PICS score is built as the average over every combination of positive and negative control.

2.5 Clustering

The final step of our pipeline aimed to identify cell populations, which is challenging given the fact that both the number of clusters and the actual cell identity are not known *a priori*. The presence of noise and the data containing large number of dimensions complicate the task too. This step is usually tackled by unsupervised clustering algorithm. In particular graph-based methods have gained a lot of interest in recent years and performed well in the context of scRNA-seq [12] as well as of CITE-seq. We started by comparing Louvain and Leiden clustering algorithms, both based on modularity to find the best clustering in a greedy fashion [4, 14]. A partition should highlight structures that could not be found in random graphs, i.e. the more structured a partition is compared to a random graph, the better the partition. The modularity of a partition is a scalar value between -1 and 1 measuring the density of intra-cluster links as compared to inter-cluster ones [4], so higher modularity means better partitioning. For weighted graphs, the modularity Q is calculated as follows:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where $A_{i,j}$ is the weight of the edge between vertices i and j, $k_i = \sum_i A_{i,j}$ is the degree of vertex i, c_i is the community to which vertex i is assigned to, δ is the Dirac function with $\delta(u, v) = 1$ if $u = v$ and 0 otherwise, lastly $m = \sum_{i,j} A_{i,j}$.

The algorithm used in Louvain considers every vertex as individual cluster at first, then picks every vertex at random order, removing it from its current cluster to add it to the cluster which leads to the highest modularity gain. This step is repeated as long as cluster membership changes, then the vertices of the resulting clusters are aggregated in a single vertex per cluster (bottom-up strategy) and the algorithm is run again as long as modularity improves.

Leiden clustering is very similar to Louvain, but runs faster and provides better results than its predecessor [14]. This algorithm introduces an additional step, in between the local moving of nodes and the aggregation of the network already found in Louvain. That additional step refines the partition in the sense that the same cluster might lead to more than one vertex after aggregation. While all those vertices will have the same label at first, the next round of local moving might find a better partition, particularly in the context of disconnected communities.

We also investigated the clustering tool provided by CiteFuse, a package consisting of a suite of tools for CITE-seq data analysis. The exciting promise of CiteFuse is to outperform previous methods developed for scRNA-seq by combining the observation of RNA and cell-surface protein available from CITE-seq experiment instead of using one or the other species [17]. CiteFuse builds a similarity network from the fusion of both the ADT and RNA similarity matrices and then performs spectral clustering on it.

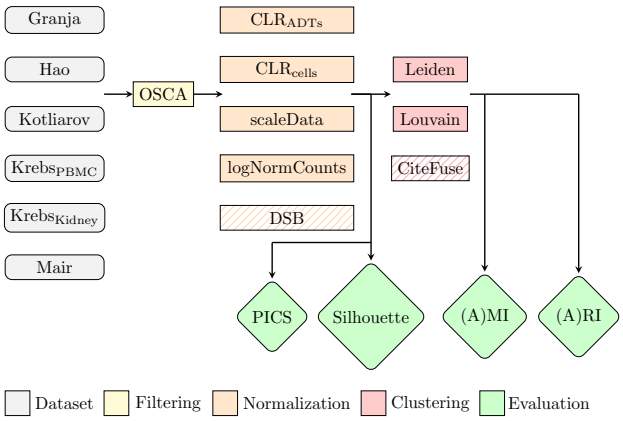


Figure 1: The diagram above depicts the different alternatives used at each step of the pipelines. First, one of the datasets was filtered, then one of the normalization techniques was applied and the normalized data was clustered. Finally, metrics are computed to evaluate the performance of the normalization or the whole pipeline. DSB normalization and CiteFuse clustering (hatched) were not included in the main analysis but represent interesting addition for further work as they were specifically designed for CITE-seq data.

2.6 pipeComp

The R package pipeComp was used to efficiently implement the combination of the different alternatives illustrated in Figure 1. This package was designed to provide a general framework for the evaluation of any computational pipeline even though it was initially applied for scRNA-seq analysis pipeline benchmark [16]. Every step can be expanded to try different sets of parameters and any metrics can easily be added to the analysis if not found among the default ones. For instance, an option was available in the code used to try different parameters influencing the resolution of the graph built for later clustering. Also, the PICS score was inserted into the analysis while the other evaluation could be found in the default tools offered by pipeComp. This package offers a way to neatly develop the foundation of benchmarks which could face subsequent integration of more datasets, functions, parameters or evaluation metrics.

3 Results

Following the steps of our pipeline, the first result we present relates to the comparison of normalization methods. Figure 2 exposes the silhouette, which led to surprisingly small values. That silhouette coefficient evaluated how well cells of a specific type defined by the ground truth were aggregated together in the reduced space after normalization on the ADT counts. A low or negative silhouette coefficient means that a cell was found to be grouped with other cells which are not similar to it, according to the ground truth used. The minimum silhouette width displays the worst case of each cell type, i.e. the coefficient of the cell which was the most wrongly arranged. The mean silhouette width

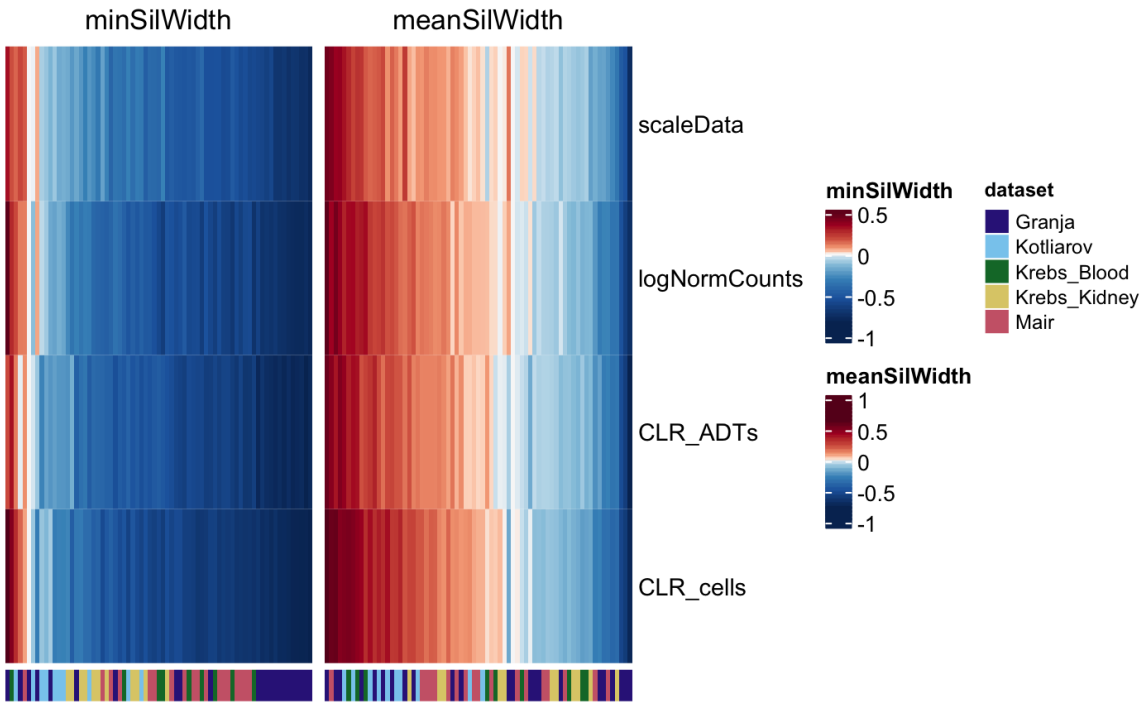


Figure 2: The heatmap above presents the silhouette of each group of cells after normalization. Each column corresponds to one cluster of a specific dataset. The higher the silhouette width, the better the agreement between the clustering and the classification defined by the ground truth. The minimum silhouette width (minSilWidth) presents the worst case of each cluster, whereas the mean silhouette width (meanSilWidth) presents the average width over all the cells of a cluster. CLR_{ADTs} uses geometric mean across ADTs whereas $\text{CLR}_{\text{cells}}$ uses geometric mean across cells.

accounts for the average over the coefficient of all cells found in one type. We see that no minimum silhouette coefficient was found to be larger than 0.5. We notice that most cell types found in Granja et al. dataset formed a block of very low minimum silhouette width. The cell types from Mair et al. also led to low width, even though not as extreme as Granja et al. Those two datasets had the worst width for some cell types, but they also got positive width for some of their types. On the other hand, all the cell types from the Krebs et al. kidney dataset obtained negative minimum silhouette width, and none of them got a mean silhouette width larger than 0.5. The dataset which earned the best widths overall was the one from Kotliarov et al.. Our results were not as high as the ones from the paper presenting pipecComp. This paper had minimum silhouette of above 0.5 for certain types, as well as a smaller proportion of negative coefficient overall [16]. The different methods gave similar outputs, with the exception of CLR_{ADTs} which produced even smaller values than its alternatives. From supplementary Figure C.1, one can see that the ground truth contained cell subtypes that were difficult to tell apart. The monocytes were divided into three different groups (CD14.Mono.1, CD14.Mono.2 and CD16.Mono) but appeared as one cluster in the UMAP visualization. The same issue was found in Figure C.5 for the monocytes, but also for the different B cells (naive and memory).

We then evaluated the performance of the different normalization on two datasets using our novel score. $\text{CLR}_{\text{cells}}$ performed better overall with respect to the PICS score, even though logNormCounts produced the largest score for the dataset from Kotliarov et al. while performing badly for data from Hao et al.. Again, it was difficult to pinpoint the very best normalization technique, as none of them clearly stood out.

Finally, two alternative clustering approaches were used and global evaluation metrics were computed. The results of the entire pipelines were gathered into the heatmap shown in Figure 4. The rand index (RI) compares two clusterings and evaluates their concordance. In our case, the clustering produced by our pipeline was compared to the ground truth. It returns values between -1 and 1, where 1 indicates a perfect agreement between the two partitions. When using the adjusted rand index (ARI), 0 corresponds to the performance of a random clustering. We computed both the RI and the ARI in order to investigate the influence of the number of clusters, since larger number of clusters might artificially boost the RI score whereas the ARI takes that number into account for its adjustment. Indeed, the chance of having two different cells properly assigned to two different clusters increases with the number of clusters, which in turn increases the RI score. We observed large differences in terms of both the ARI and RI across datasets but rather small ones across normalizations. A slight improvement appeared when applying the Leiden clustering algorithm instead of Louvain, notably for the Granja et al. dataset. We saw that all the indices decreased when switching to

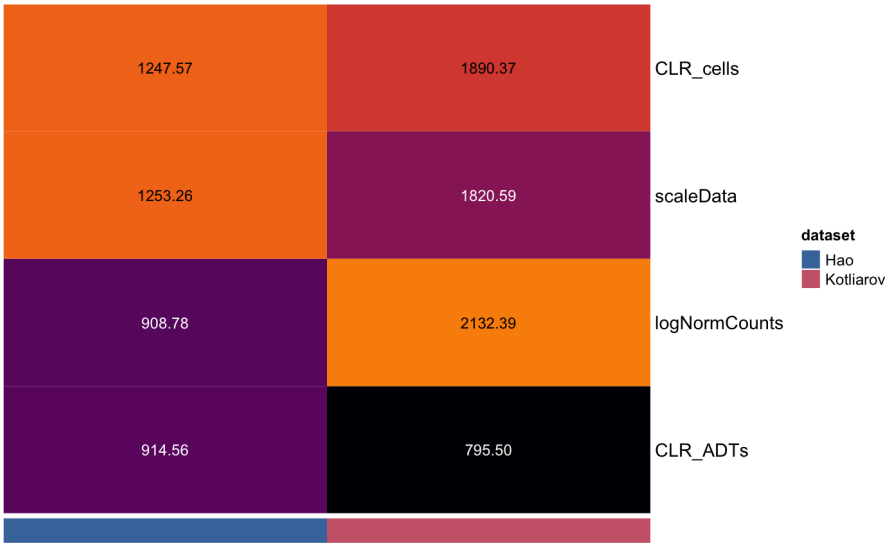


Figure 3: The heatmap above presents the PICS score for the two datasets which contained isotypes. Methods like logNormCounts were very impacted by the dataset used, whereas $\text{CLR}_{\text{cells}}$ gave similar scores for both. CLR_{ADTs} gave very poor scores compared to the other methods. Note that the actual metric values are printed, while the colors are mapped to signed square-root of the number of (matrix-wise) Median Absolute Deviations from the (column-wise) median.

the adjusted metrics, which was expected since the ARI takes the RI relative to an expected value [2]. The dataset from Kotliarov et al. appeared to be the least impacted by the adjustment. The best combination was logNormCounts followed by Leiden clustering applied to Kotliarov et al., but the MI shown in complementary Figure D.1 favored $\text{CLR}_{\text{cells}}$ rather than logNormCounts. The AMI gave the same ranking than ARI when the Louvain clustering was used, but CLR_{ADTs} gave better AMI than scaleData in combination with Leiden clustering.

Comparing the results from the silhouette width, the ARI and the PICS score, we get a blurry picture of what happened. Both CLR alternatives appear as the worst normalization in terms of silhouette and ARI, but $\text{CLR}_{\text{cells}}$ still performed well relative to the PICS score.

4 Discussion

In the previous section, we saw that the silhouette widths were not as high as expected, that it was difficult to clearly decide which normalization method performed the best and that the difference between the alternative pipelines scores seemed to arise mostly from the choice of dataset and its respective ground truth. Since the silhouette was generally low for every normalization, we believe that even though it failed at telling which method was best, it still indicated the source of the problem: the low silhouette could have been caused either by a poor discrimination of the different cell types based on the low dimensional representation of the cells given their ADT profile or by a mismatch between the clustering resolution of the ground truth and that same mapping of cells on a reduced space. It is also probable that the problem was caused by a combination of both those suppositions. If the first hypothesis is true, then more ADTs would be needed to properly discriminate the different cell types. If the second hypothesis is true, then more attention should be given to the way the ground truth used for the evaluation was set. After inspecting the mapping for the combination of every dataset and normalization shown in supplementary Figures C.1, C.2, C.3, C.4 and C.5, we concluded that good cell partition was achieved when a large set of ADTs was available and that the ground truth described general cell types. For instance, the dataset from Kotliarov et al. surely benefited from its 87 ADTs to infer the 9 cell types. On the other hand, datasets like the one from Granja et al. was undermined by a smaller number of ADTs to deduce affiliation of a cell to one of the 26 very precise types. Since the Mair et al. and Granja et al. datasets still led to positive width for some cell types, we hypothesis that we should lower our expectations for certain cell subtypes by considering only the general type like monocytes to have meaningful analysis of the methods. On the other hand, the Krebs et al. kidney dataset must be the victim of a mismatch between the ADTs used and the cell types that are to be identified, i.e. the ADTs might have been usefull in combination of the RNA profile of the cell in their study, but they don't allow to identify precise cell types alone. A similar reasoning about the ground truth might explain the low score of data from Mair et al. despite its larger number of ADTs. Also, there might have been a mismatch between the 42 ADTs present in Mair et al. and the ones that would have been needed to effectively distinguish the cell types. Lastly, the two datasets from Krebs et al. focused on CD3^+ T cells, which sub-types might not be possible to discriminate without relying on transcriptomics.

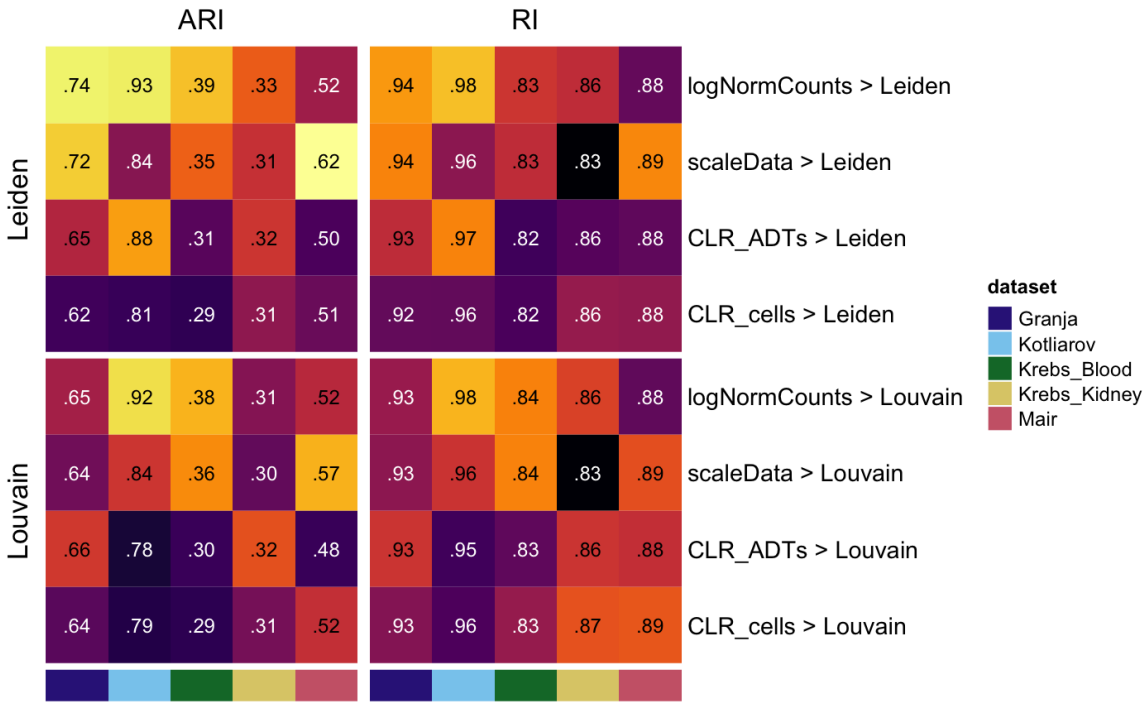


Figure 4: The heatmap above presents the (adjusted) rand index, which evaluates the similarity between the clusterings performed through our pipelines and the one used by the authors providing the datasets. Each row corresponds to a different pipeline defined by the choice of normalization and clustering method, with > indicating the order in which the steps are computed. The rows are ordered with respect to the overall performance in terms of ARI. Note that the actual metric values are printed, while the colors are mapped to signed square-root of the number of (matrix-wise) Median Absolute Deviations from the (column-wise) median.

The PICS score did not help much to decide which normalization would be best to use, but it placed CLR_{ADTs} at the bottom of the ranking and that correlated to some extent with the silhouette results. A surprising outcome was the large difference between logNormCounts score for data from Hao et al. and Kotliarov et al.. This observation incited us to have a closer look at the selected positive control. As shown in supplementary Figures B.1 and B.3, it appeared that the choice of positive controls was not properly done, with the exception of CD45-2 in the Hao et al. dataset. Indeed, the mode of those raw counts distribution was found to stand below 20, and even below 10 for CD45-1 in Hao et al., which does not correspond to what was expected of a positive control. This behavior can be explained by the fact that even though CD45 is expressed on every leukocytes, it is actually expressed in different isoforms and all those isoforms might not be expressed simultaneously in all leukocytes. For instance, regulatory T cells express either CD45RA or CD45RO after activation [5], but certainly not both. Additional explorations of the markers distribution showed that other ADTs could have been used as positive controls, for instance CD93 and CD48 for Hao et al. dataset and CD18 and HLA-ABC for Kotliarov et al. dataset. The biological implications of those markers have to be investigated further.

Similarly to what was discussed on the silhouette, no strong conclusion can be drawn from the variation in terms of (A)RI with respect to the different pipeline used. The complexity of the resulting low dimensional representation was carried on from the previous step and it was expected that datasets with better silhouette would get larger metrics. The slight improvement of Leiden clustering over Louvain observed was comforting, but more effort should be invested to determine if that improvement was significant. Since the performance reported by metrics like RI is influenced by the actual number of clusters used [3], we took a look at the difference between the number of clusters predicted by our pipeline and the one defined by the ground truth. As shown in complementary Figure D.2, the dataset giving out the best metrics was also the one for which the number of cluster was largely over-estimated. Indeed, our pipelines predicted from three to eight additional clusters in the dataset from Kotliarov et al., which might have helped boost the metrics. It is not clear why this dataset was the least impacted by the adjustment, with the largest gap between RI and ARI being 0.17 for the pipeline using either CLR_{ADTs} or $\text{CLR}_{\text{cells}}$ with Louvain clustering. On the other hand, our pipeline under-estimated the number of clusters to be found in data from Granja et al., but the large number found in the ground truth might have helped that dataset to get good metrics too. We remind that, even though the number of cluster might have had some influence, the Kotliarov et al. dataset was shown earlier to lead to the best low dimensional representation and

it make sens that, being easier to cluster, it got a better score.

5 Conclusion

In conclusion, we brought together diverse datasets to assess the attainable quality of cell type clustering from their ADT counts only. We showed that a satisfying cell partition was achievable as long as datasets with enough (distinctive) ADTs were available and that the resolution aimed went along with the ADTs at hand. In our case, most datasets were challenging to cluster on due to their high cell-type resolution and small ADT panel. We used a variety of methods to score alternative approaches, and we laid the foundation of a new ground truth independent scoring function. Even though most pipeline alternatives were difficult to rank, we would advise not using CLR_{ADTs} for normalization.

6 Further work

This report describes the first step towards a more comprehensive use of ADT information in single cell analysis. Even though some interesting matters have been uncovered, there is still an immense amount of work to be done. We would recommend to start by refining the ground truth to be more in line with the resolution we can hope to achieved using the available ADTs. That way, good or bad performance could be linked with the chosen method instead of the dataset used. Next, one should use the positive control ADTs proposed or look more into which alternatives exist for the PICS score. It came to our attention that a clusterability measure called SIGMA [21] could be turned into an ground truth independent score, offering an alternative to PICS. Techniques specially designed for CITE-seq data were left out, for the reason that the data at hand were not compatible (e.g. for DSB) or that the techniques did not fit in the time or memory constraints (e.g. for CiteFuse). It would be beneficial if such normalization like DSB could be tested, by providing the necessary raw (unfiltered) ADTs counts for all datasets. Also, CiteFuse offers filtering and clustering methods taking advantage of the combination of both the RNA and ADT counts.

References

1. Altin, J. G. & Sloan, E. K. The role of CD45 and CD45-associated molecules in T cell activation. *Immunology and Cell Biology* **75,** 430–445. ISSN: 08189641. https://pubmed.ncbi.nlm.nih.gov/9429890/ (1997).
2. Yeung, K. Y. & Ruzzo, W. L. Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data" (to appear in Bioinformatics). *Science* **17** (2001).
3. Wagner, S. & Wagner, D. Comparing Clusterings-An Overview. https://i11www.iti.kit.edu/extra/publications/ww-cco-06.pdf (Jan. 2007).
4. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment,* P10008. ISSN: 17425468. https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008 (2008).
5. Booth, N. J. *et al.* Different Proliferative Potential and Migratory Characteristics of Human CD4 + Regulatory T Cells That Express either CD45RA or CD45RO. *The Journal of Immunology* **184,** 4317–4326. ISSN: 0022-1767. http://www.jimmunol.org/content/184/8/4317http://www.jimmunol.org/content/184/8/4317.full#ref-list-1 (Apr. 2010).
6. Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nature Publishing Group.* https://www.nature.com/articles/nmeth.2769.pdf?origin=ppub (2014).
7. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2,** 15. ISSN: 20492618. http://www.microbiomejournal.com/content/2/1/15http://www.microbiomejournal.com/content/2/1/15 (2014).
8. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33,** 495–502. ISSN: 15461696. https://www.nature.com/articles/nbt.3192 (2015).
9. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9.** https://genomemedicine.biomedcentral.com/track/pdf/10.1186/s13073-017-0467-4.pdf (2017).
10. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology* **35,** 936–939. ISSN: 15461696 (2017).
11. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14,** 865–868. ISSN: 15487105. https://www.nature.com/articles/nmeth.4380 (2017).
12. Freytag, S., Tian, L., Lönnstedt, I., Ng, M. & Bahlo, M. Comparison of clustering tools in R for medium-sized 10x genomics single-cell RNA-sequencing data [version 1; referees: 1 approved, 2 approved with reservations]. *F1000Research* **7,** 1297. ISSN: 1759796X. https://doi.org/10.12688/f1000research.15809.1 (2018).

13. Bushel, P. R. *et al.* Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. https://doi.org/10.3389/fgene.2018.00636 (2019).

14. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9.** ISSN: 20452322. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6435756/ (2019).

15. Amezquita, R. A. *et al.* Data infrastructure Orchestrating single-cell analysis with Bioconductor. *Nature Methods* **17.** https://github.com/Bioconductor/OrchestratingSingleCellAnalysis (2020).

16. Germain, P. L., Sonrel, A. & Robinson, M. D. PipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. *Genome Biology* **21,** 227. ISSN: 1474760X. https://doi.org/10.1186/s13059-020-02136-7 (2020).

17. Kim, H. J., Lin, Y., Geddes, T. A., Yang, J. Y. H. & Yang, P. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics* **36,** 4137–4143. ISSN: 14602059. https://academic.oup.com/bioinformatics/article/36/14/4137/5827474 (2020).

18. Mulè, M., Martins, A. & Tsang, J. Normalizing and denoising protein expression data from droplet-based single cell profiling. *bioRxiv,* 2020.02.24.963603. https://doi.org/10.1101/2020.02.24.963603 (2020).

19. Stewart, C. A. *et al.* Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nature Cancer* **1,** 423–436. https://pubmed.ncbi.nlm.nih.gov/33521652/ (2020).

20. Germain, P.-L. scDblFinder. *Bioconductor.* https://github.com/plger/scDblFinder (2021).

21. Mircea, M. *et al.* A clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations. *bioRxiv.* https://doi.org/10.1101/2021.05.11.443685 (2021).

Supplement

A Code

The code used for this project is available on the following GitHub repository:
https://github.com/johaab/CITEnREAPseq

B Isotypes and Positive controls

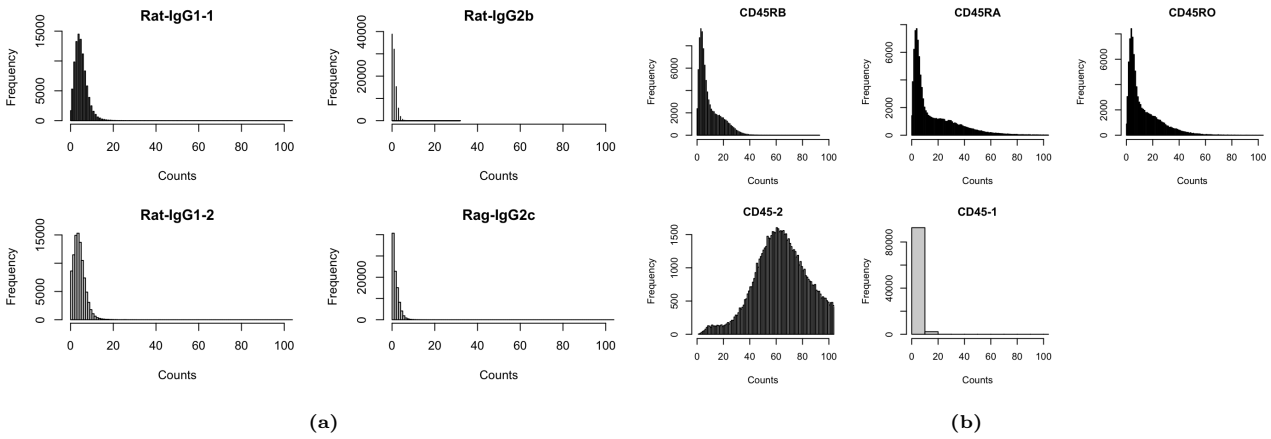


Figure B.1: Raw counts distribution of the ADTs from Hao et al. dataset chosen as isotypes **(a)** and positive controls **(b)**. We see that even though the isotypes distribution followed what was expected, only one of the positive controls (CD45-2) did. Hence, poor PICS score could have simply be caused by erroneous positive controls and more effort should be made for the careful selection of the ADTs used as positive control.

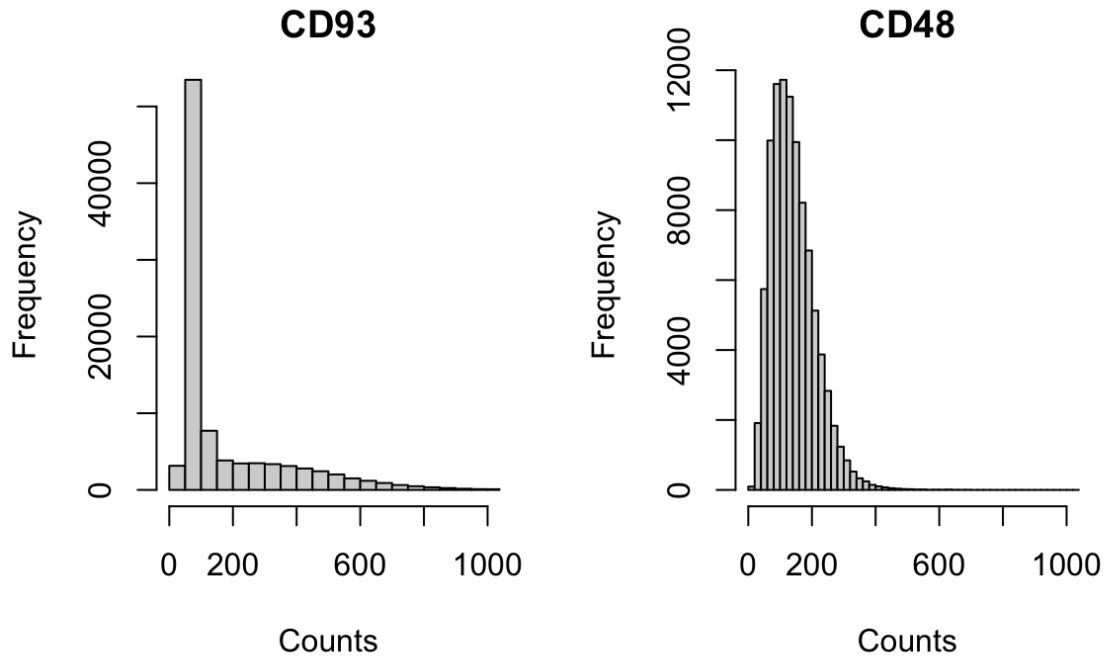
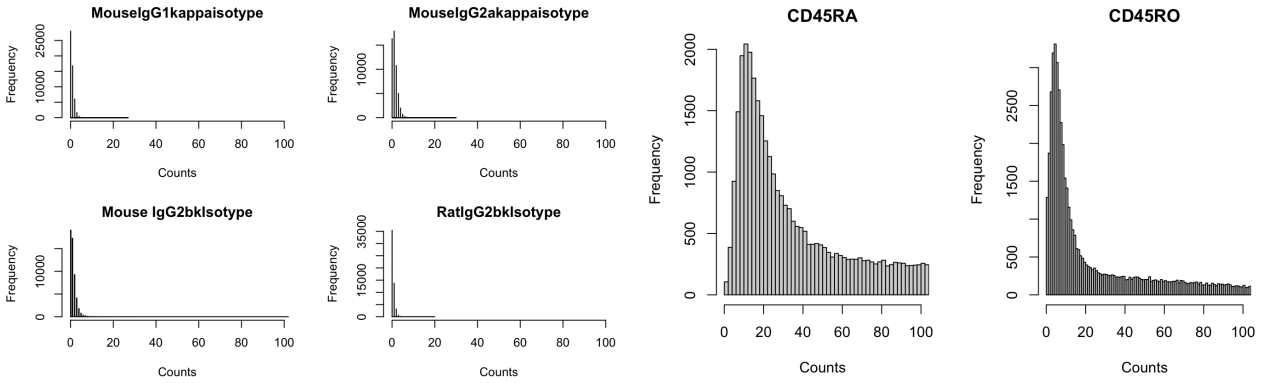


Figure B.2: Raw counts distribution of the ADTs from Hao et al. dataset that could be used as positive controls in further work. The mode of the CD93 and CD48 counts distribution is 93 and 105 respectively, both larger than any of the modes shown in Figure B.1. The biological meaning of those marker has yet to be checked.



(a) **(b)**

Figure B.3: Raw counts distribution of the ADTs from Kotliarov et al. dataset chosen as isotypes **(a)** and positive controls **(b)**. We see that even though the isotypes distribution follow what was expected, it is not the case for the positive controls. The positive control distributions have long right tail indicating that some cells expressed such markers to a high level but the fact that the mode is so low implies that most cells did not. Hence, more effort should be made on the careful selection of the ADTs used as positive control.

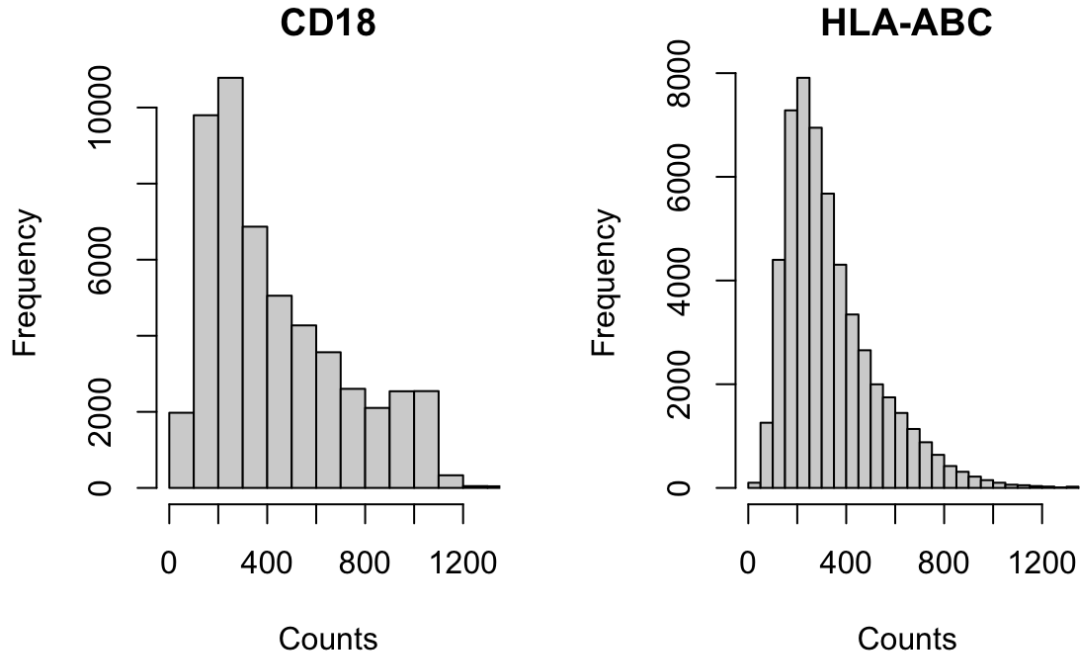


Figure B.4: Raw counts distribution of the ADTs from Kotliarov et al. dataset that could be used as positive controls in further work. The mode of the CD18 and HLA-ABC counts distribution is 193 and 183 respectively, both larger than any of the modes shown in Figure B.3. The biological meaning of those marker has yet to be checked.

C Low dimensional representation

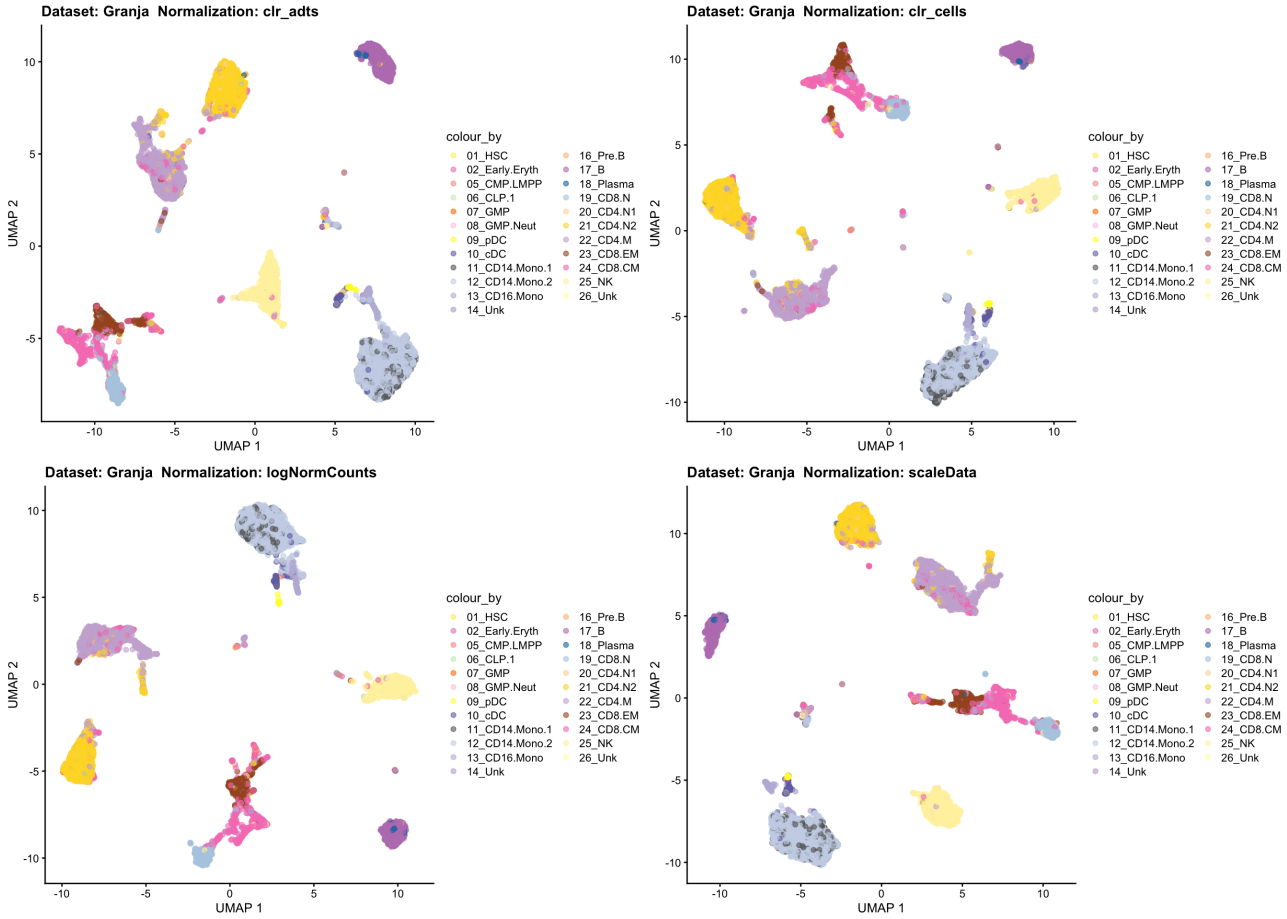


Figure C.1: Low dimensional representation of Granja et al. dataset after different normalization methods have been applied to the ADT counts, coloured by the authors annotation. Clear groupings are observed but did not agree entirely with the fine grained labels from the ground truth used.

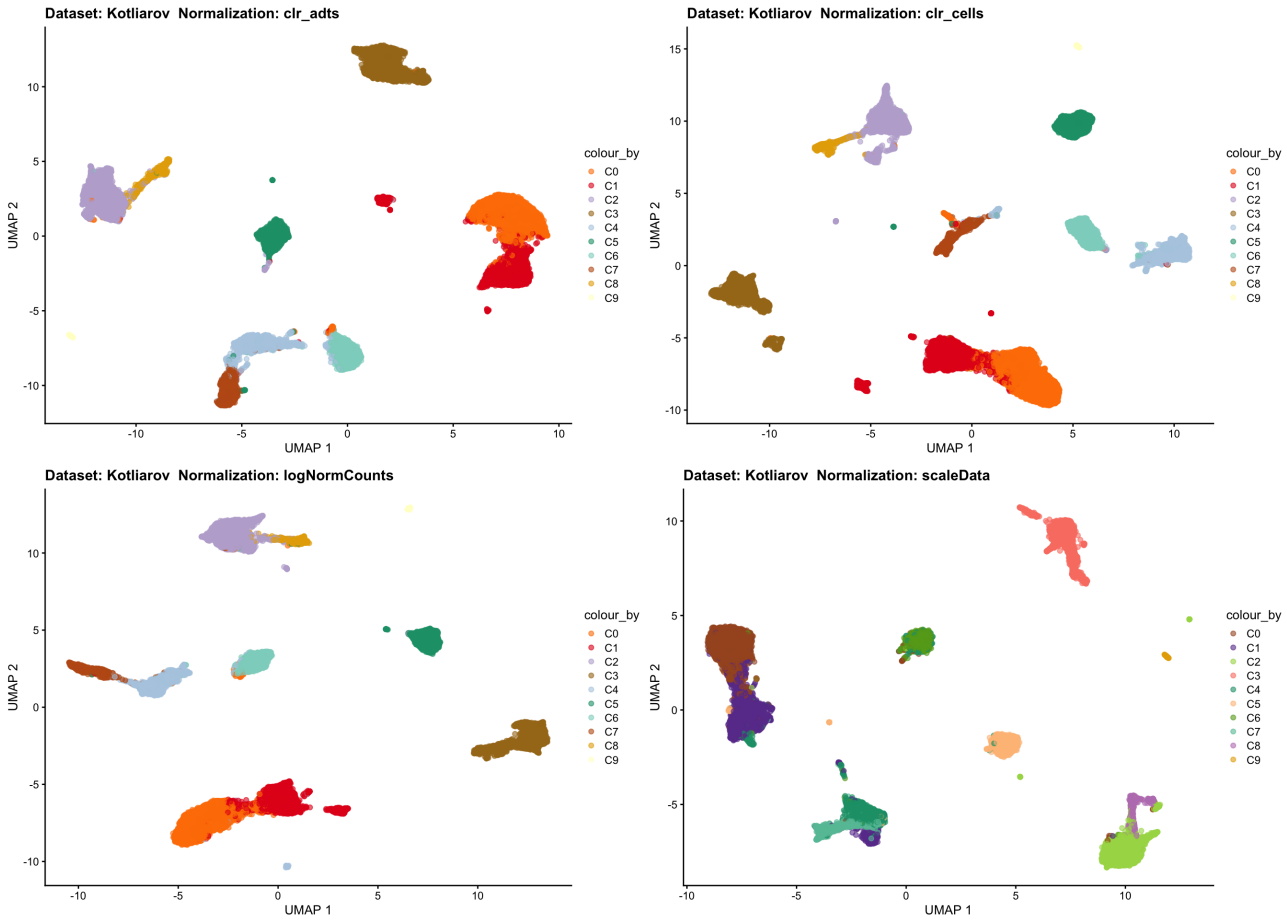


Figure C.2: Low dimensional representation of Kotliarov et al. dataset after different normalization methods have been applied to the ADT counts, coloured by the authors annotation. The cell types are well segregated in this reduced space and more or less in line with the labels stated in the ground truth.

C0: $CD4^+$ naive T/DNT
C1: $CD4^+$ memory T
C2: Classical monocytes and mDC
C3: B cells
C4: $CD8^+$
C5: NK cells
C6: $CD8^+$ naive T
C7: Unconventional T cells
C8: Non-classical monocytes
C9: pDC

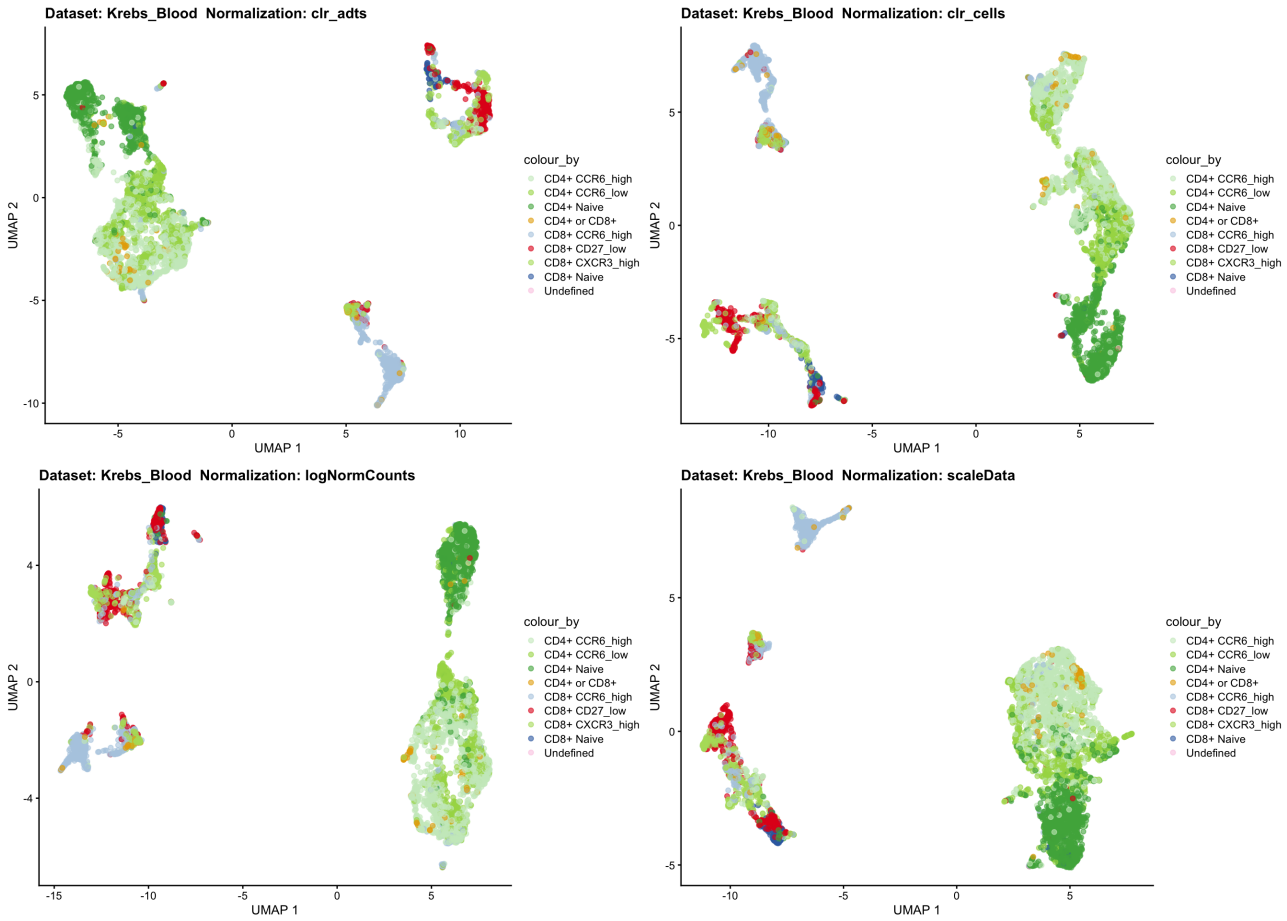


Figure C.3: Low dimensional representation of Krebs et al. dataset (PBMCs) after different normalization methods have been applied to the ADT counts, coloured by the authors annotation. We see that the ADTs did not manage to separate the cell types to the level of detail carried by the ground truth. Having only two cell types (e.g. $CD4^+$ and $CD8^+$) seems more reasonable.

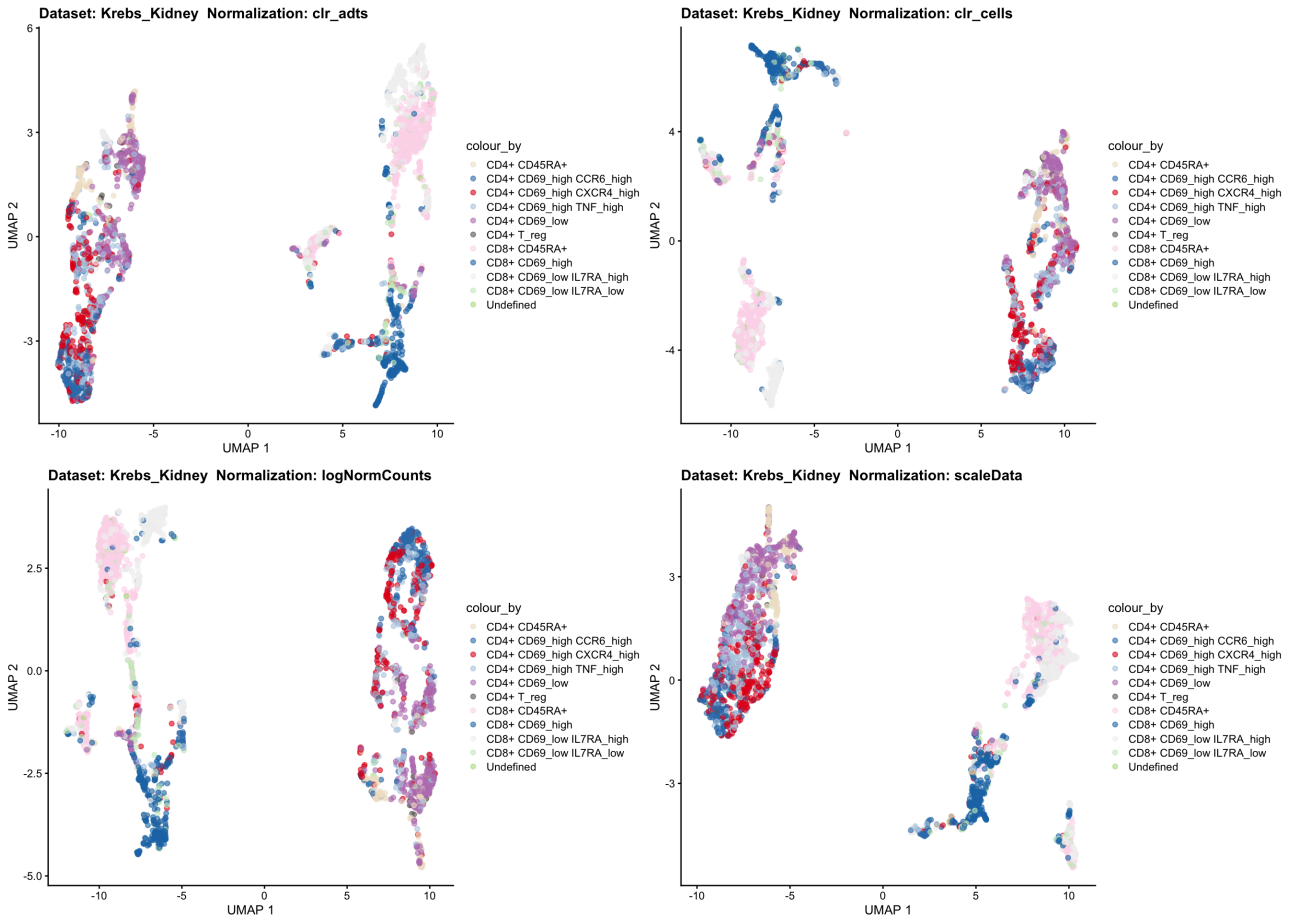


Figure C.4: Low dimensional representation of Krebs et al. dataset (Kidney) after different normalization methods have been applied to the ADT counts, coloured by the authors annotation. We see that the ADTs did not manage to separate the cell types to the level of detail carried by the ground truth. Having only two cell types (e.g. $CD4^+$ and $CD8^+$) seems more reasonable.

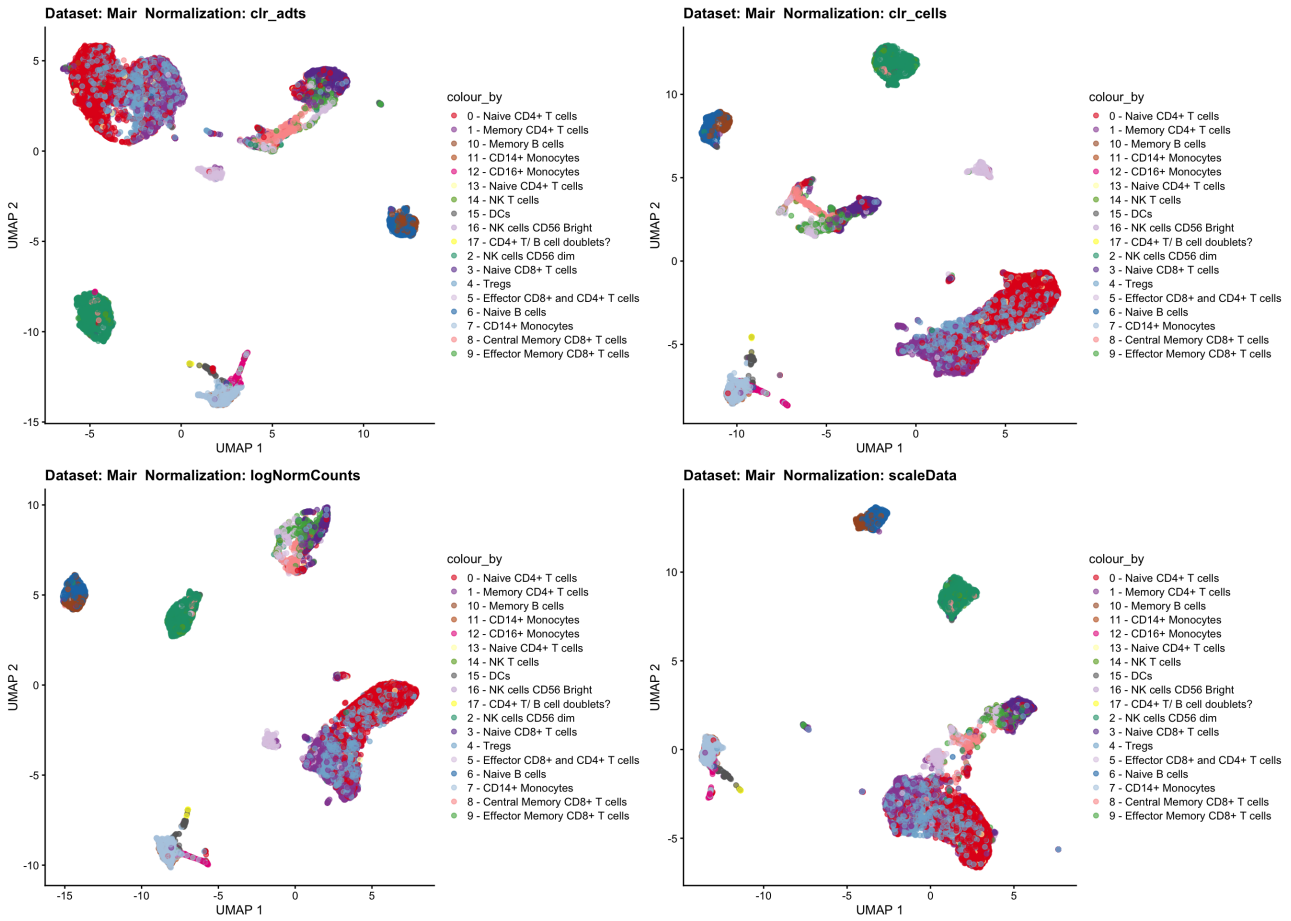


Figure C.5: Low dimensional representation of Mair et al. dataset after different normalization methods have been applied to the ADT counts, coloured by the authors annotation. The ADTs managed to extract groups of cells but the ground truth was too refine to agree completely with the clusters.

D Other

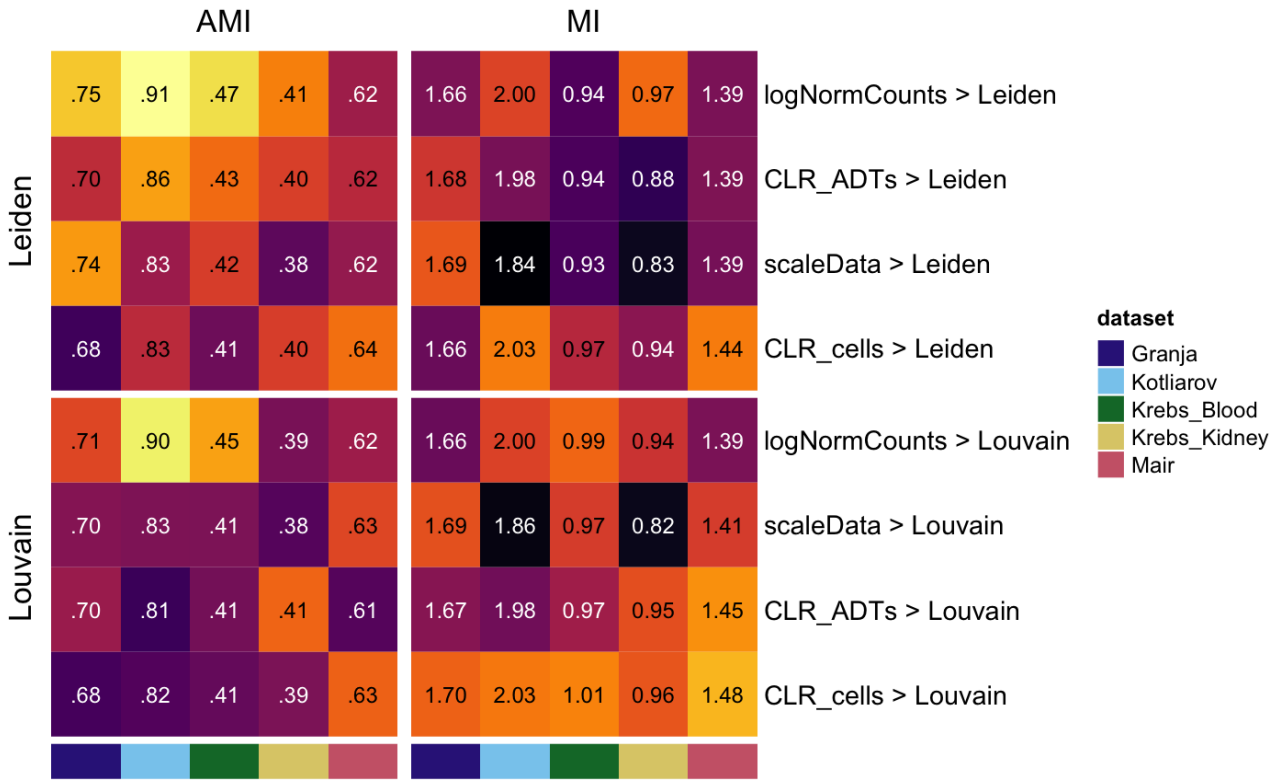


Figure D.1: The heatmap above presents the (adjusted) mutual information, which evaluates the similarity between the clusterings performed through our pipelines and the one used by the authors providing the datasets. Each row corresponds to a different pipeline defined by the choice of normalization and clustering method, with > indicating the order in which the steps are computed. The rows are ordered with respect to the overall performance in terms of AMI. Note that the actual metric values are printed, while the colors are mapped to signed square-root of the number of (matrix-wise) Median Absolute Deviations from the (column-wise) median.

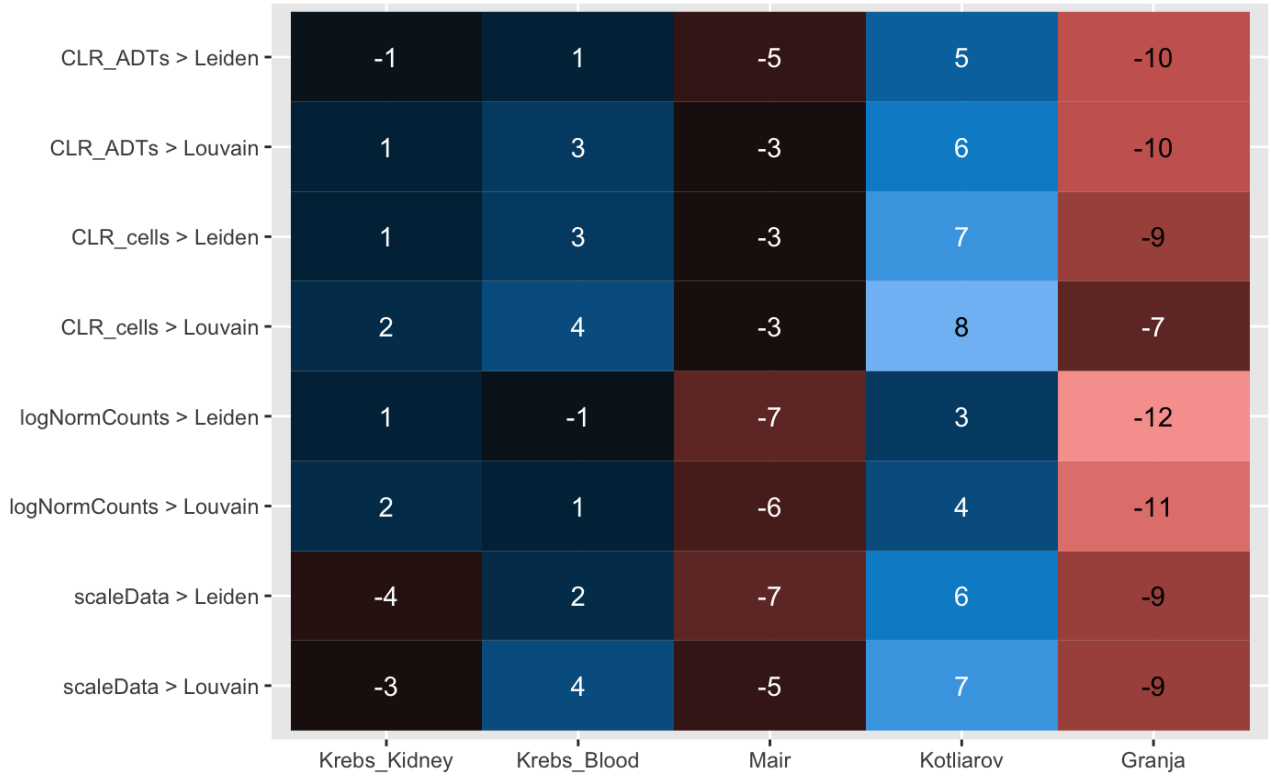


Figure D.2: The heatmap above presents the difference in terms of number of clusters between each pipeline alternative and the ground truth. Each row corresponds to a different pipeline defined by the choice of normalization and clustering method, with > indicating the order in which the steps are computed. The columns are ordered with respect to the absolute difference. The coloring is done with respect to the difference observed, with black meaning no or small difference, light blue (red) meaning strong positive (negative) difference. Positive (negative) difference indicates that the pipeline gave more (less) clusters than the ground truth contained. Our pipelines appeared to over- and under-estimate the number of clusters in the datasets from Kotliarov et al. and Granja et al., respectively.