732A76 Research Project

Identifying cell types within a single cell data biological analysis. Comparison between ScCATCH (Marker Gene Method) and CIPR (Correlation Based Method)

Author: Carolina Rosário
Supervisor: Oleg Sysoev
Linköping University.

Abstract

Single cell data biological analysis includes a primary step that is investigating which type each cell belongs to CIPR (Correlation based method) and ScCATCH (Marker gene method) methods were analysed and numerically compared under different situations.

Introduction

This project has the objective of supporting single cell data biological analysis by investigating one important initial step of the process, that is, determine which type each cell belongs to. To make such annotation, various machine learning methods were proposed [Pasquini] and the most representative method from each of the two following categories was chosen for this research. From the category "Marker gene method" the chosen method was ScCATCH and from the category "Correlation based method" the chosen method was CIPR. A detailed numerical comparison of these methods was made in order to conclude which method obtained better results, under the following situations:

- **1:** The method contains exactly same cell types as cell types in the single cell data.
- **2:** The method contains more cell types than in single cell data.
- **3:** Some cell types that are in single cell data are not present in the method.
- **4:** There are various proportions of cell types in single cell data.

Methods

After some research regarding which methods produced better outcomes among the mentioned automated methods for cell type annotation on scRNA-seq data [Pasquini], CIPR and ScCATCH were chosen.

Correlation Based Method: CIPR

CIPR (Cluster Identity PRedictor), a web-based R/Shiny app and R package, provides a graphical user interface to quickly score gene expression profiles of unknown cell clusters against mouse or human references, or a custom dataset provided by the user. Against other functionally similar software methods, CIPR is less computationally demanding, it performs significantly faster and provides accurate predictions [Ekiz].

Marker gene method: ScCATCH

In ScCATCH (Single-cell Cluster-based Automatic Annotation Toolkit for Cellular Heterogeneity), another R package, cell types are annotated through the tissue-specific cellular taxonomy reference database (CellMatch) and the evidence-based scoring (ES) protocol. ScCATCH facilitates analysis on scRNA-seq data and outperformed Seurat, a popular method for marker genes identification, and cell-based annotation methods [Shao].

Results

The dataset used for this study includes 12500 human peripheral blood cells and 17938 genes. There are 5 cell types, CD4 (T cells), CD8 (T cells), CD14 (Monocytes), CD19 (B cells), and CD56 (Natural Killer – NK cells), containing 2500 cells each. (Figure 1).

Main-Cell	Sub-Cell	Count of Cells
T Cells	CD4	2500
	CD8	2500
Monocytes	CD14	2500
B Cells	CD19	2500
NK Cells	CD56	2500

Figure 1

A graphical representation of the data is presented below using UMAP, a dimensional reduction technique, from Seurat tool (Figure 2).

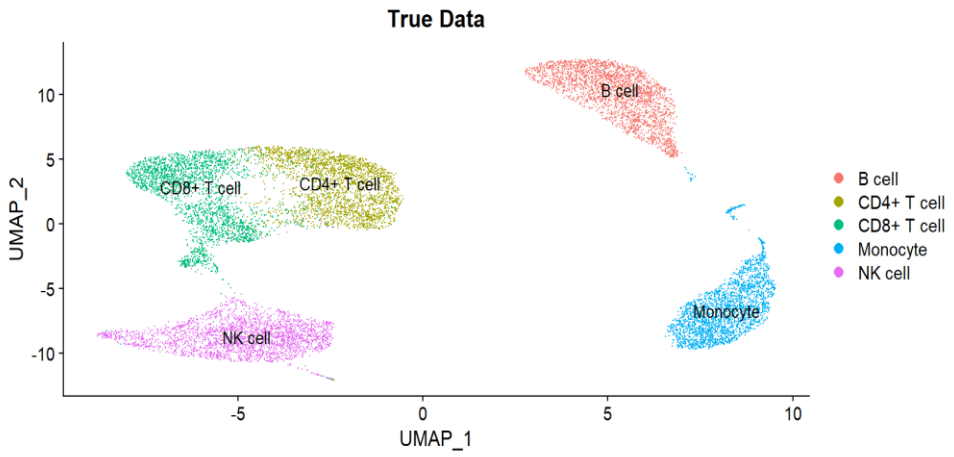


Figure 2

Pre-Processing of Data

Originally, the used data had genes in the form of ensembled gene ID's. During pre-processing, these were converted to gene symbols since both CIPR and ScCATCH had the need of receiving the genes in the gene symbol form. To do this conversion, different experiences were made with different tools, such as *biomaRt* and *annotables* packages in R and *Mygene* in python. The number of effective converted genes was used to support the decision, and *Mygene* performed best when compared with the other tools, converting successfully 15743 out of 17398 genes (approx. 88%). Therefore, one of the initial steps taken in this research project was to pre-process the data in python using *Mygene*. The pre-processed data was then saved in python_pre_processed_5celltypes_2500each.txt and loaded later in R to perform the analysis. A small part of the original data and its conversion is showed below, as an example (Figure 3).

Ensembled Gene ID	CD4_Th__cell_1	CD4_Th__cell_2	…
ENSG00000130695	0	0	…
ENSG00000142669	3	3	…
ENSG00000158062	1	0	…
ENSG00000169442	5	2	…
ENSG00000176092	0	0	…
…	…	…	…

Gene Symbol	CD4_Th__cell_1	CD4_Th__cell_2	…
CEP85	0	0	…
SH3BGRL3	3	3	…
UBXN11	1	0	…
CD52	5	2	…
CRYBG2	0	0	…
…	…	…	…

Figure 3

CIPR and ScCATCH Analysis

Both algorithms were tested under the 4 situations described in the beginning of the report. Additionally, the same analysis was made for a smaller number of clusters and a larger number of clusters to understand whether it could have some impact in the performance of the chosen algorithms.

Some common steps were taken for both algorithms, starting with a standard pre-processing workflow for scRNA-seq data in Seurat, including data normalization and scaling, and the detection of highly variable features. Then, PCA is performed to reduce dimensionality and identify possible clusters. When computing the clusters, the resolution parameter could be freely adjusted to find different combinations of cluster sizes. For this research, the default value was applied (0.8) and a smaller value (0.05) to get the ground truth of the number of classes. 13 and 5 clusters were generated, respectively (Figure 4 and 5). These combinations of clusters will be used for Situation 1,2 and 3. However, for situation 4, a subset of the full data will be used for the analysis in order get to get various proportions of cell types in single cell data. The subset used for Situation 4 has randomly selected 2000 T cells (1000 CD4 and 1000 CD8), 500 NK cells, 100 B cells and 800 Monocytes. Same resolution values were applied, and 11 and 4 clusters were generated (Figure 6 and 7).

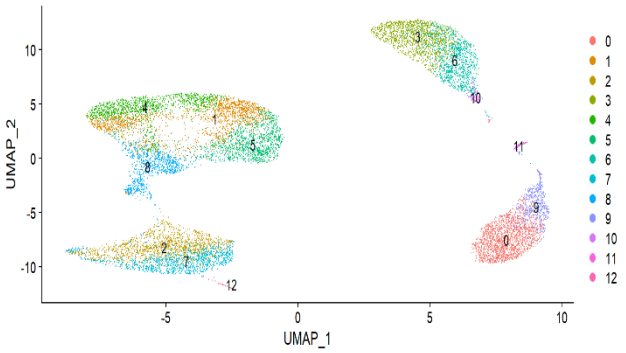


Figure 4 – General Case - 13 Clusters

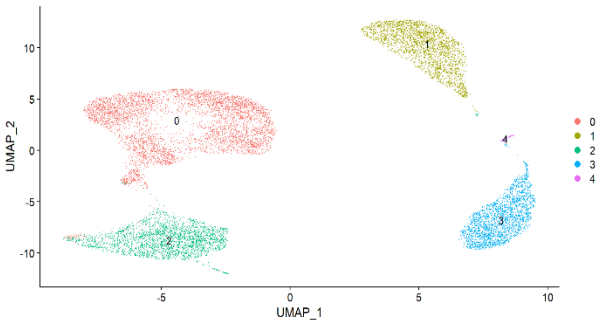


Figure 5 – General Case - 5 Clusters

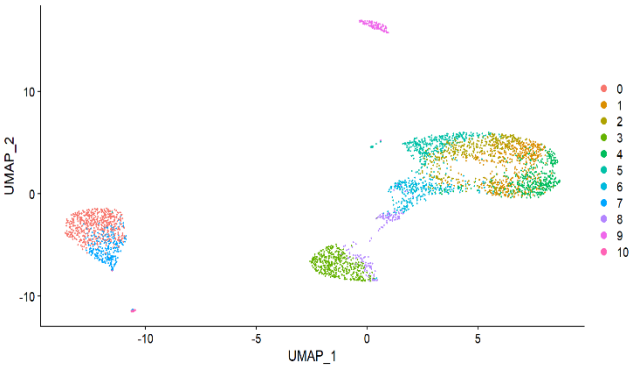


Figure 6 – Subset Case - 10 Clusters

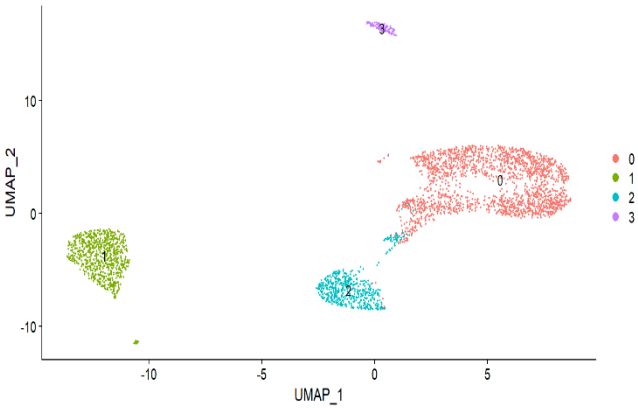


Figure 7 – Subset Case - 4 Clusters

Cluster Classification Results

Situation 1: The method contains exactly same cell types as cell types in the single cell data. The dataset contains T cells (CD4 and CD8), NK cells, B cells and Monocytes.

Therefore, the method was filtered to contain the same cell types (Figure 8 – Situation 1 Results).

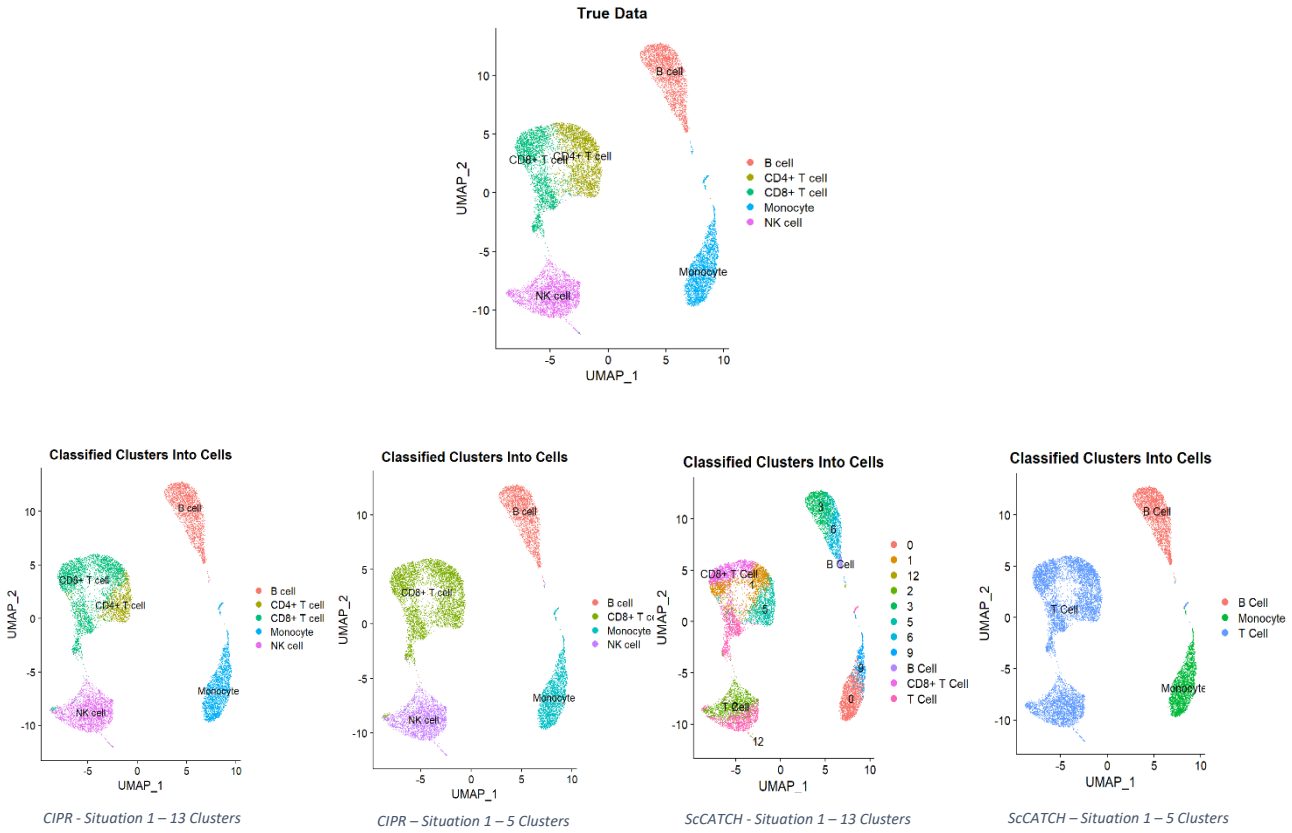


Figure 8 – Situation 1 Results – True Data and Classified Clusters

Situation 2: The method contains more cell types than in single cell data. All available cells in the method were used (Figure 9 – Situation 2 Results).

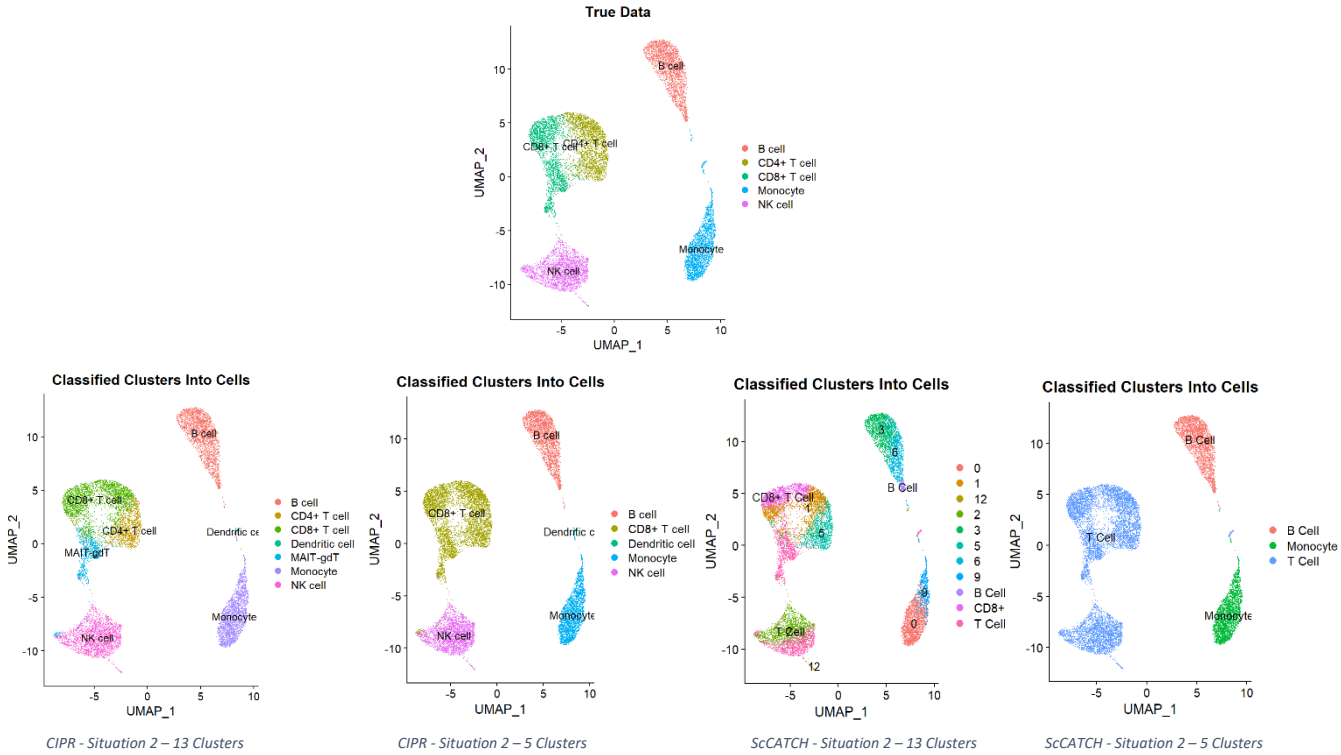
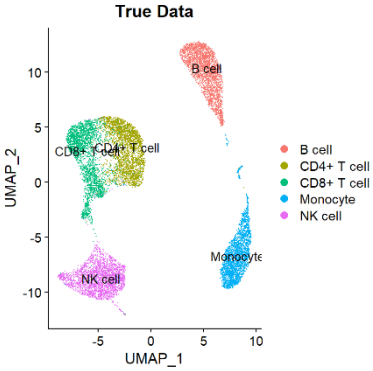
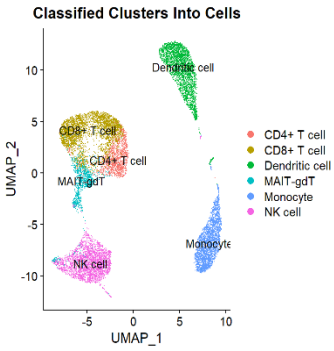


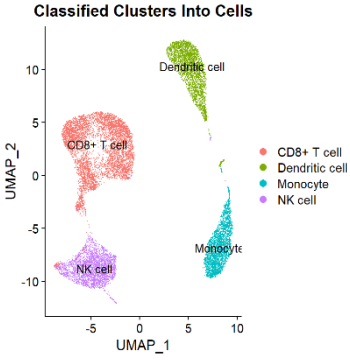
Figure 9 – Situation 2 Results – True Data and Classified Clusters

Situation 3: Some cell types that are in single cell data are not present in the method. In this case, B cells are not present in the method, even though they exist in single cell data.

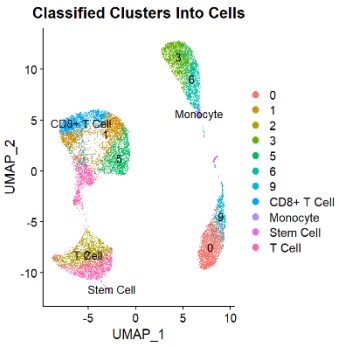




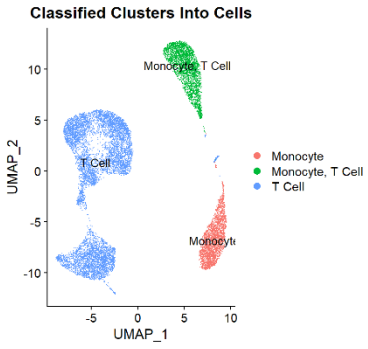
CIPR - Situation 3 – 13 Clusters



CIPR - Situation 3 – 5 Clusters



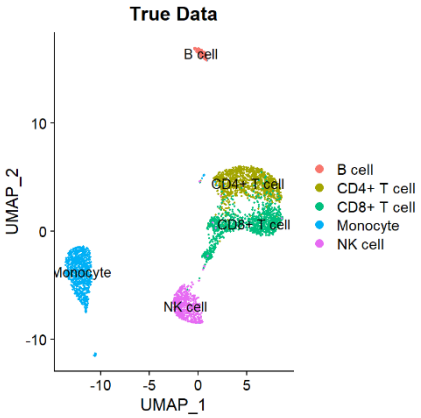
ScCATCH - Situation 3 – 13 Clusters

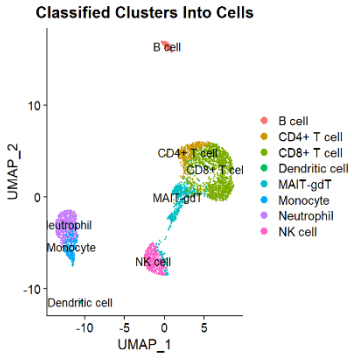


ScCATCH - Situation 3 – 5 Clusters

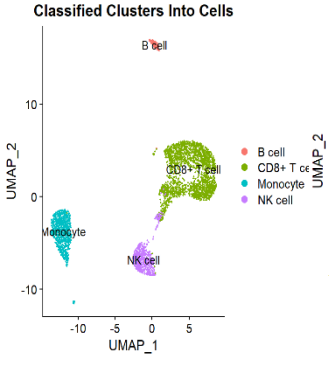
Figure 10 – Situation 3 Results

Situation 4: There are various proportions of cell types in single cell data. Single cell data subset with randomly selected 2000 T cells (1000 CD4 and 1000 CD8), 500 NK cells, 100 B cells and 800 Monocytes was used for the analysis.

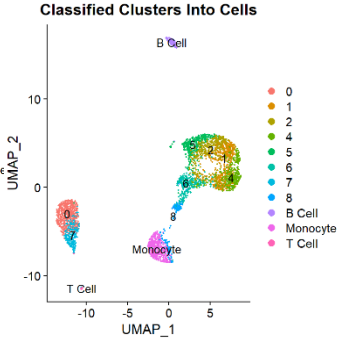




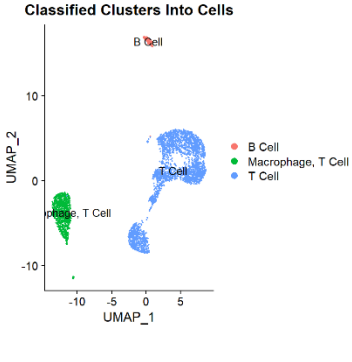
CIPR - Situation 4 – 10 Clusters



CIPR - Situation 4 – 4 Clusters



ScCATCH - Situation 4 – 10 Clusters



ScCATCH - Situation 4 – 4 Clusters

Figure 11 – Situation 4 Results

Accuracy

During the analysis, for each tested situation, various measures were calculated including accuracy, that was calculated as:

$$Accuracy = \frac{Number\ of\ correctly\ classified\ cells}{Total\ number\ of\ cells}$$

This measure will be used as a measure of performance to evaluate whether either Correlation Based Methods or Marker Gene Methods have a better performance, having into account various situations and cluster sizes. For each situation, best accuracy is highlighted in green and the worst in red.

	Description	**Correlation based method CIPR**		**Marker gene method ScCATCH**	
		Larger number of clusters	**Smaller number of clusters**	**Larger number of clusters**	**Smaller number of clusters**
Situation 1	The method contains exactly same cell types as cell types in the single cell data.	87%	79%	18%	79%
Situation 2	The method contains more cell types than in single cell data.	80%	78%	18%	79%
Situation 3	Some cell types that are in single cell data are not present in the method.	60%	58%	17%	59%
Situation 4	There are various proportions of cell types in single cell data.	50%	68%	3%	62%

Figure 12 – Accuracy Results

Discussion

Based on the full analysis and accuracy results (Figure 12), some conclusions can be taken. Overall, CIPR didn't show any significant differences in performance regarding the number of clusters used for analysis. However, ScCATCH significantly improved its accuracy when the number of clusters is closer to the true number of classes, showing a very poor performance when the number of clusters is, in this example, higher.

As expected, of all situations, *situation 1* have shown best results, since the existent cell labels in the methods are limited to the true cell labels and therefore, the probability of error while classifying the clusters will be lower.

Situation 2 also achieved good results, quite similar to *situation 1* results, which proves that, both methods can make good predictions even when a broader group of cells is used for the classification.
In *situation 3*, the goal was to understand how both methods would classify a certain cell group, in this case B cells, when that cell is not present in the method. CIPR classified B cells as Dendritic cells for both cluster groups. In ScCATCH, for a smaller number of clusters, only a very small part was classified at all, and it was as Monocytes. For a larger number of clusters, the same method classified B cells as T cells and Monocytes.

In situation 4, it's being evaluated how the methods perform when the classes are imbalanced, that means, when there are various proportions of cell types in single cell data, and for that, a subset of the original data was created with different numbers of samples per class, as described before. Even though the mean accuracy of all tests is approximately 60%, some cells were very well classified, even when there's a very low proportion of those cells in the data when compared to other cells. For example, there were only 100 B cells in the new dataset, and all of them were correctly classified in all methods. Additionally, Natural Killer cells were also well classified in CIPR, but poorly in ScCATCH.

It's possible to conclude that, in this research, correlation based method had a better performance than marker gene method. The correlation based method (CIPR) did a better prediction of the cell label, regardless of the number of clusters used for analysis and that the number of clusters used for analysis has a high impact in the performance the marker gene method (ScCATCH).

References

[Pasquini] Pasquini, G., Arias, J. E. R., Schäfer, P., & Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. Computational and Structural Biotechnology Journal, 19, 961-969.

[Ekiz] Ekiz, H.A., Conley, C.J., Stephens, W.Z. et al. CIPR: a web-based R/shiny app and R package to annotate cell clusters in single cell RNA sequencing experiments. BMC Bioinformatics 21, 191 (2020)

[Shao] Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., & Fan, X. (2020). scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. In iScience (Vol. 23, Issue 3, p. 100882). Elsevier BV.