# Analysis of sc-RNA-seq of colorectal cancer

**Maria Bochenek**

**colorectal cancer, Seurat, clustering, doublets**

## Indroduction

The immunotherapy for colorectal cancer is effective only for certain types of tumors. thus there has been extensive multiomics studies on molecular landscape of colorectal cancer (CRC) that allowed the classification of cells from colorectal cancer to consensus molecular subtypes (CMS). Characterisation of tumor complexity using single-cell transcriptome analysis can help to dissect molecular landscape by cellular component, which can offer a deeper understanding of tumor heterogeneity. Presence of certain types of cells like for example cancer-associated fibroblasts, which are the hallmarks of the mesenchymal phenotype CMS4 are associated with poor survival rate. (1)

Many current anticancer therapies target nontumor components such as extracellular matrix, immune system and vascular system. Thus, identifying cellular components of the tumor landscape is vital to choosing viable therapy.

In this article we analyze single-cell RNA sequencing (sc-RNA-seq) data to try to understand and classify cellular components of colorectal cancer. We we conduct our analysis according to current best practices of working with sc-RNA-seq data. (2).

## Methods

**The dataset.** The raw scRNA-seq data analyzed in this article comes from a study conducted by the Molecular Digestive Oncology Unit Department of Oncology in Katholieke Universiteit Leuven in 2018. The dataset named KUL3 was later used in a bigger study (Lee et al). Samples come from 6 Belgian patients diagnosed with CRC who underwent surgery without previous treatment. After surgical resection samples from tumor core, tumor border, and adjacent non-malignant colon tissue from the same resection specimen were collected and immediately transferred for tissue preparation. Half of the tissues were subjected to single cell isolation and RNA-seq of coding RNA. Fresh single-cell suspensions were loaded into the Chromium system (10x Genomics) targeting 5,000 cells. Sequencing data were aligned to the human reference genome (GRCh38) and processed using CellRanger 2.1.0 pipeline (10x Genomics). Further sequencing protocol information as well as raw data are available in the ArrayExpress under the ascention code E-MTAB-8410 [1].

The dataset downloaded from Single Cell Expression Atlas contained 52609 cells. In the analysis we used only samples assigned to 6 adult individuals with subsequential id KUL01, KUL19, KUL21, KUL28, KUL30, KUL31, which consisted of 34484 cells. Further information about samples can be found in Figure 1.

[1] https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8410/

| | Sample | Sample size | Individual | Sex | Age (years) | Organism part | Sampling site |
|---|---|---|---|---|---|---|---|
| 1 | Sample1c | 2992 | KUL01 | female | 81 | caecum | normal tissue |
| 2 | Sample1b | 3328 | KUL01 | female | 81 | caecum | tumour border |
| 3 | Sample1a | 2495 | KUL01 | female | 81 | caecum | tumour core |
| 4 | Sample2c | 4015 | KUL19 | female | 86 | rectosigmoid junction | normal tissue |
| 5 | Sample2b | 3920 | KUL19 | female | 86 | rectosigmoid junction | tumour border |
| 6 | Sample2a | 2253 | KUL19 | female | 86 | rectosigmoid junction | tumour core |
| 7 | Sample3c | 2644 | KUL21 | female | 50 | sigmoid colon | normal tissue |
| 8 | Sample3b | 2209 | KUL21 | female | 50 | sigmoid colon | tumour border |
| 9 | Sample3a | 1705 | KUL21 | female | 50 | sigmoid colon | tumour core |
| 10 | Sample6c | 590 | KUL28 | male | 52 | sigmoid colon | normal tissue |
| 11 | Sample6b | 225 | KUL28 | male | 52 | sigmoid colon | tumour border |
| 12 | Sample6a | 1149 | KUL28 | male | 52 | sigmoid colon | tumour core |
| 13 | Sample8c | 315 | KUL30 | male | 84 | ascending colon | normal tissue |
| 14 | Sample8b | 386 | KUL30 | male | 84 | ascending colon | tumour border |
| 15 | Sample8a | 2537 | KUL30 | male | 84 | ascending colon | tumour core |
| 16 | Sample9c | 360 | KUL31 | male | 85 | sigmoid colon | normal tissue |
| 17 | Sample9b | 1506 | KUL31 | male | 85 | sigmoid colon | tumour border |
| 18 | Sample9a | 1855 | KUL31 | male | 85 | sigmoid colon | tumour core |

**Fig. 1.** Experiment design and information about samples

**Filtering the raw gene expression matrix.** We analyzed the raw gene expression matrix from the CellRanger pipeline using the Seurat R package which turned out to be superior to the RCAv2 R package which couldn't manage to cluster cells using inbuilt functions.

The raw gene expression matrix was loaded and used to create the Seurat object. Then the expression matrix has been filtered following the criteria stated in Lee et al. The cells with more than 1000 unique molecular identifiers (nCount_RNA > 1000), more than 200 and less than 6000 genes (nFeature_RNA > 200 and nFeature_RNA < 6000) and less than 20% of mitochondrial gene expression in UMI counts.

Then we performed data normalization using the LogNormalize method implemented in NormalizeData function in Seurat, which normalizes the feature expression measurements for each cell by the total expression, then multiplies it by the scale factor (we used the default 10 000) and finally log-transforms the result. (1)

**Feature selection.** The next step was finding the subset of features that exhibit high cell-to-cell variation in the data in order to only focus on those highly variable genes in further analysis. Research shows that this type of feature selection can help to amplify biological signals in single-cell datasets. The feature selection was done using an algorithm implemented in FindVariableFeatures function in Seurat package with following parameters: selection method as "vst", return 2000 features, feature mean values between 0.0125 and 3 and feature dispersion greater than 3.

Top 10 most variable features

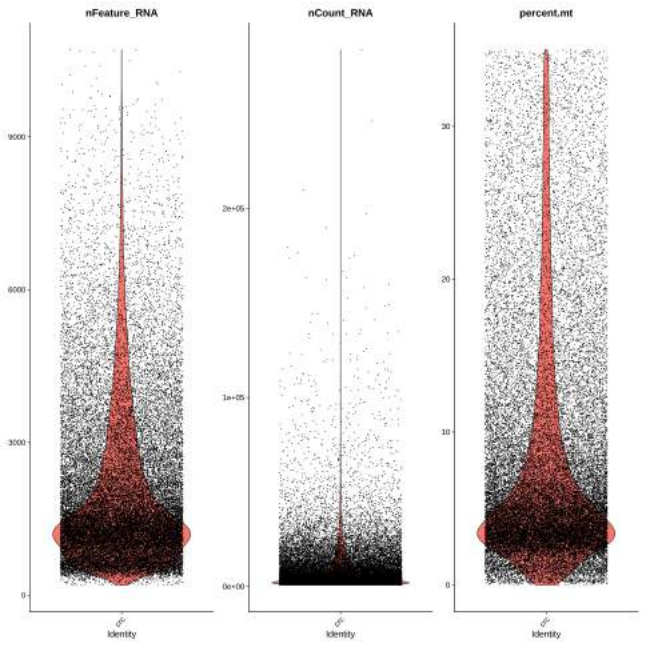


**Fig. 2.** the QC metrics

1. ENSG00000197253 (*TPSB2*) - Tryptase Beta 2
2. ENSG00000254709 (*IGLL5*) - Immunoglobulin Lambda Like Polypeptide 5
3. ENSG00000253755 (*IGHGP*) - Immunoglobulin Heavy Constant Gamma P (Non-Functional)
4. ENSG00000282094 (*IGHGP*) - Immunoglobulin Heavy Constant Gamma P (Non-Functional)
5. ENSG00000174992 (*ZG16*) - Zymogen Granule Protein 16
6. ENSG00000282184 (*IGHG3*) - Immunoglobulin Heavy Constant Gamma 3 (G3m Marker)
7. ENSG00000274497 (*IGHG2*) - Immunoglobulin Heavy Constant Gamma 2 (G2m Marker)
8. ENSG00000277016 (*IGHG4*) - Immunoglobulin Heavy Constant Gamma 4 (G4m Marker)
9. ENSG00000172236 (*TPSAB1*) - Tryptase Alpha/Beta 1
10. ENSG00000211897 (*IGHG3*) - Immunoglobulin Heavy Constant Gamma 3 (G3m Marker)

We can observe that genes coding constant regions of immunoglobulin heavy chains (*IGHG2, IGHG3, IGHG4*) as well as other immunoglobulin (*IGHGP, IGLL5*) have varying levels of expression in the cells. This is expected as we are dealing with cancer and normal cells which are characterized by different levels of immunoglobulin expression (usually higher in cancer cells) (3). The other gene with differing expression levels between cells is tryptase (*TPSB2, TPSAB1*) which is a major neutral protease in mast cells and is secreted

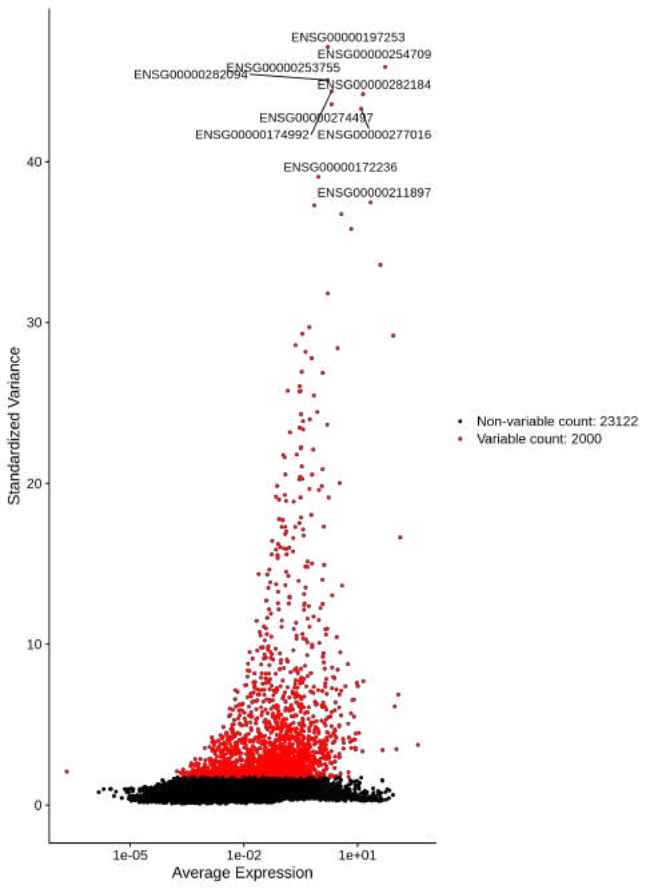


**Fig. 3.** Most variable features

upon activation-degranulation response of these cells. Mast cells accumulate in the tumors and their microenvironment during disease progression (4) so varying levels of tryptase gene expression between cancer and normal cells is expected. The ZG16 gene is associated with mucus-secreting cells, especially Intestinal goblet cells and is one of the most significantly down-regulated genes in colorectal cancer tissues (5).

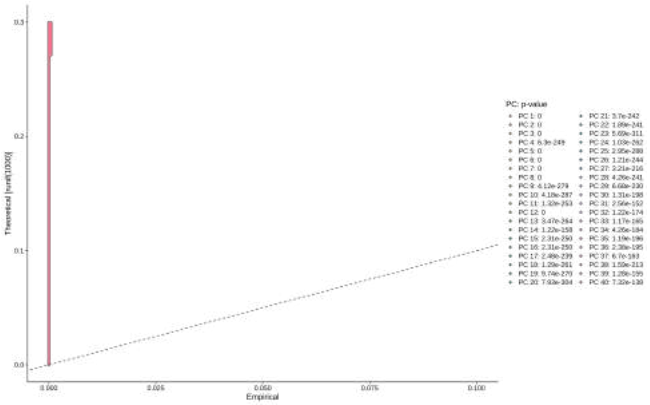

**Fig. 4.** Significance of top 40 PCs

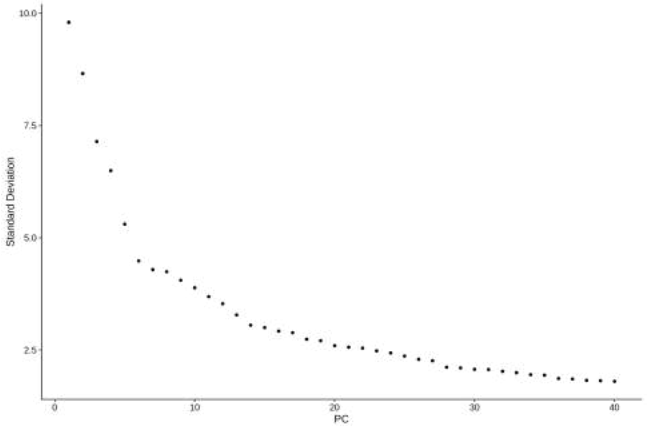


**Fig. 5.** Elbow plot of top 40 PCs

**Dimentional reduction.** Next step in pre-processing our data is scaling prior to dimensional reduction techniques. We use the algorithm implemented in the ScaleData function from the Seurat package. This step assigns equal weight to the genes so that in downstream analysis the highly expressed ones do not completely dominate the signal. After that the data is ready for dimension reduction techniques. We perform PCA using selected highly variable features.

To determine the "dimensionality" of the dataset we test which PCs are statistically significant using Jackstraw and ScoreJackstaw functions with the first 40 PCs.

However as seen on Figure 4 p-values of all 40 PCs are significant and very close to 0 which suggests that high number of PCs significantly explain the variance of the dataset, thus reducing dimensionality by choosing the few top PCs might result in loss of significant biological signaling reflected in the data.

After analyzing the Elbow plot Figure 5 we decided to choose the first 28 PCs for further analysis. The choice was substantiated by further clustering simulations using a greater number of dimensions which resulted in a high number of clusters. This high of a number was unnecessary as assigning cell types to clusters is a difficult problem that requires prior knowledge about cell markers and gene expression in different cell types. Thus we decided to choose a lower number of dimensions to focus on more global expression patterns in the dataset rather than trying to capture more delicate differences.

| Multiplet Rate (%) | # of Cells Loaded | # of Cells Recovered |
|---|---|---|
| ~0.4% | ~825 | ~500 |
| ~0.8% | ~1,650 | ~1,000 |
| ~1.6% | ~3,300 | ~2,000 |
| ~2.4% | ~4,950 | ~3,000 |
| ~3.2% | ~6,600 | ~4,000 |
| ~4.0% | ~8,250 | ~5,000 |
| ~4.8% | ~9,900 | ~6,000 |
| ~5.6% | ~11,550 | ~7,000 |
| ~6.4% | ~13,200 | ~8,000 |
| ~7.2% | ~14,850 | ~9,000 |
| ~8.0% | ~16,500 | ~10,000 |

**Fig. 6.** Multiplet rate is linearly dependent on the amount of loaded cells

**Removing the doublets.** Next step in preprocessing is finding and removing doublets, which are barcodes where more than one cell ends up in a droplet. Overloading of cells is a common problem in scRNA-seq protocols in droplet-based methods. In a typical 10x experiment proportion of doublets is linearly dependent on the amount of loaded cells. Our data comes from an experiment using the Chromium system targeting 5000 cells. So the expected number of doublets is 682 cells according to the Chromium user guide Figure 6.

We used the DoubletFinder R package. As we don't have prior knowledge which $pK$ value to choose, thus we use paramSweep_v3 function with 28 first PCs to estimate the best $pK$ value using BC metric (Figure 7). The $pK$ value with the highest BS score is 0.005. We run doubletFinder_v3 function with 28 first PCs, $pN = 0.25$, $pK = 0.005$, $nExp = 682$. The results can be seen on Figure 8.

We can also observe (Figure 9) that predicted doublets have more detected genes than a singlets, which is to be expected. Often only top 10 PCs are used, however with our high dimensional data results after only using 10 PCs ($pK$ is then estimated as 0.17). Then doublets are found mostly in one neighborhood (coincidentally corresponding to cancer cells and the same few samples), which is less likely compared to results when using 28 PCs when doublets are found in many clusters, coming from many samples Figure 10.

After removing doublets we are left with 27366 cells.

**Influence of sample number on variance.** After analyzing PCA, UMAP and tSNE plots we can observe that the variance observed in the dataset might be explained by sample number. As each sample corresponds to certain individual with certain sex and age as well as to organism part and sampling site we can also note some grouping when colouring by each of these features. In order to minimize the effect that samples have on variance in the data we scale the data and regress sample variable against each gene/feature. That

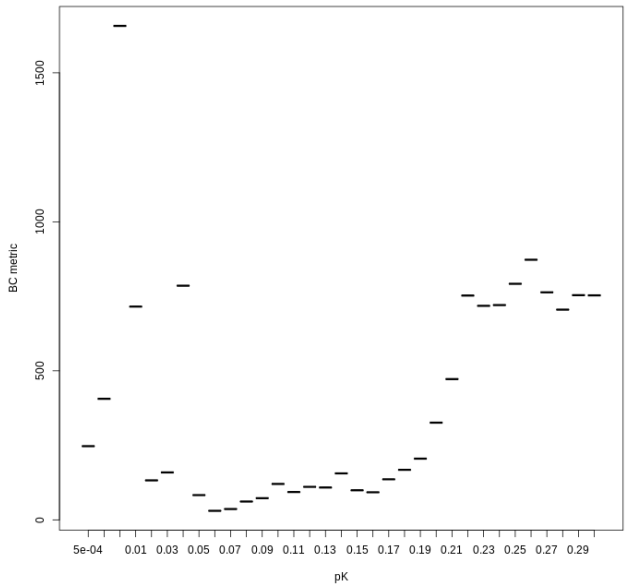


**Fig. 7.** BC metric score of pK estimations done by paramSweep_v3 function from DoubletFinder

hopefully will allow us to get more accurate clustering results.

**Unsupervised clustering.** We conducted unsupervised clustering on data scaled with regressing out for samples and without correction both times using Luovain algorithm implemented in FindClusters() function.
Resolution 0.6, 0.8, 1, 1.2 were explored for deciding on optimal clustering both before (15) and after (17) correction. Furthermore we checked how cluster cells were distributed between clusters depending on resolution before (16) and after (18) correction.
Finally have set the resolution parameter to 0.6 in both cases.

When clustering the data not corrected for the sample the algorithm found 26 clusters (plot and plot) and after correction 44 clusters.

Furthermore we searched for cluster biomarkers: differentially expressed features in different clusters. We used FindAllMarkers function to find positive markers for every cluster compared to all remaining cells. We chose to only test genes that are expressed in 0.1 cells in either of the two populations and ones that show on average, at least X-fold difference (log-scale) between the two groups of cells (logfc.threshold= 0.25).

Finally we analyze cluster markers and try to make an inference about the cell types. We used both approaches first: analyze top markers by cluster and based on them try to inferre cell type and second: identify clusters by searching where specific cell type markers are highly expressed.
Moreover we used the sctype_score function from ScType R package which allows us to automatically assign cell identity to clusters and is compatible with Seurat package (6). The sctype_score function was called only for positive cluster markers.

## Results

**Correcting possible batch effects.** After analyzing PCA, tSNE and UMAP plots we can easily notice that cells with the same metadata variable are often grouped together. However here only information of possible batches we have are samples which don't include information about possible differences coming from the experiment environment. Even though we may be eager to correct data by regressing out for each metadata variable it may not necessarily be useful in this case. When studying cancer it has been noted that variables such as sex and age might influence survival rate of certain types of cancer (CITE). Thus if we regress out those variables it may obscure important biological variables that could be useful during making inference. Similarly other metadata variables like sampling site and organism part actually bring important information in understanding the variance in data. In conclusion the only variable that we decided to make correction for is the sample number as it indirectly encapsulates access variation possibly brought by metadata variables. However this correction would still preserve enough differences between populations grouped by metadata variables that making inference based on them would be informative. This can be observed on plots before and after correction coloured by sex, age and organism part.
To summarize, in the situation where we don't have the obvious batches to correct for like different protocols or equipment, correcting for metadata variables depends on the type of inference we plan to make and the biological process itself.

**Assigning identity to clusters.** While analyzing clustering results we used two different approaches. First one is to analyze top markers by the cluster and try to inferre cell type that way. Second approach is to try to identify cluster types by searching where specific cell type markers are highly expressed. In this approach really useful is Human Protein Atlas (7) [2].

***Clustering before correction.*** For each cluster we identified markers that differentiate it from other clusters. Most of the time analysis of most significant markers for each cluster should give us some idea what kind of cell we should expect in the cluster. Top ten most significant markers for each clusters can be seen in the 27.
All clusters that we were able to assign cell identities to can be seen in Table **??**.

***Identity by cell type markers.*** The NK cell markers: NKG7 (ENSG00000105374), GNLY (ENSG00000115523) were highly expressed in cluster 3 as seen on 47. Differential cluster 3 markers included CCL4 and CCL5 which are major HIV-suppressive factors produced by CD8+ T-cells. That would suggest that cluster 3 is rather not homogenous.

Along with NK cell markers cluster 3 is characterized by the expression of $\gamma\delta$ T cells: TRDC (ENSG00000211829), TRGC2 (ENSG00000227191) and TRGC1 (ENSG00000211689) (29).

[2] https://www.proteinatlas.org

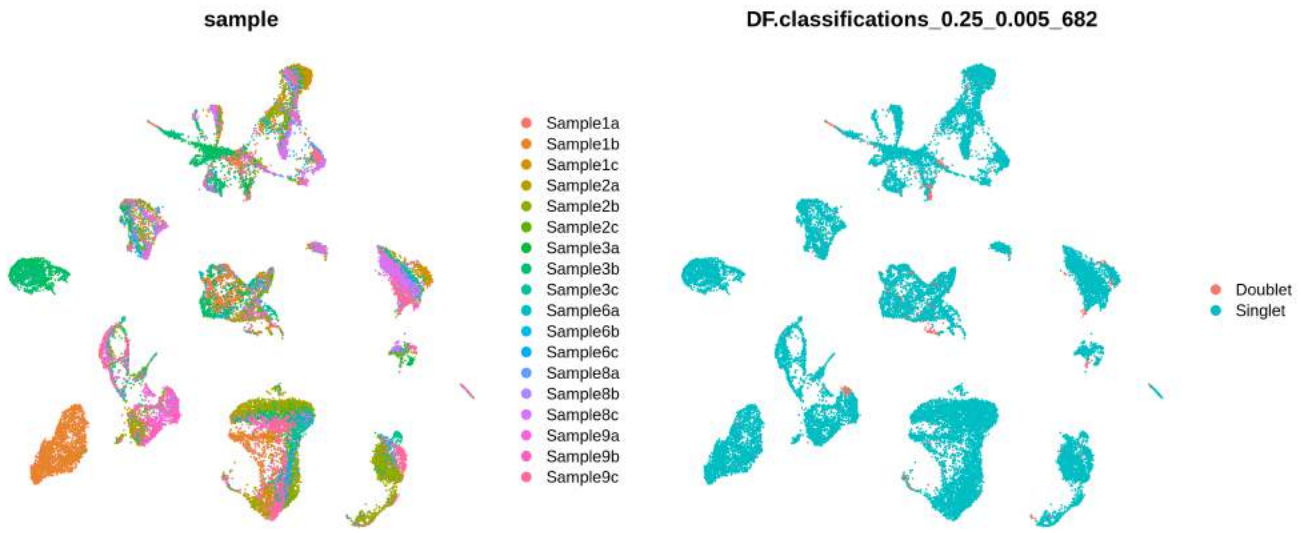


**Fig. 8.** Distribution of doublets between samples, $pK = 0.005$

| Cell type | Clusters |
|---|---|
| NK cells | 3 |
| $\gamma\delta$ T | 3 |
| CD8+ T | 3 |
| Naive CD4+ T | 0, 25 |
| Memory CD4+ T | 15 |
| CD14+ Monocytes | 9 |
| CD16+ Monocytes | 7 |
| Mast cells | 23 |
| B cells | 4, 18 |
| Neuronal cells | 19 |
| Fibroblasts | 8, 10, 12, 21 |
| IgA+ plasma cell | 1 |
| IgG+ memory B cells | 20 |
| Smooth muscle cell | 24 |

**Table 1.** Suggested cluster identity by high expression of cell type markers

Another cell population present in cluster 3 are CD8+ T cells, which were identified by their markers: CD8A (ENSG00000153563), CD8B (ENSG00000172116), which can be seen on 30.

The CD4+ T cells markers: CD4 (ENSG00000010610), CCR7 (ENSG00000126353), IL7R (ENSG00000168685) allowed us to identify naive CD4+ T (CCR7+) population in clusters 0, 25 and memory CD4+ T (CCR7-) population in cluster 15 (31).

Another cell population identified by it's canonical markers : MS4A1 (ENSG00000156738), CD19 (ENSG00000177455) were B cells found in clusters 4 and 18 as seen on 32.

Mast cells are a type of granulocyte that is part of immune and neuroimmune systems. After analysis of mast cell markers such as The mast cell markers such as TPSAB1 (ENSG00000172236), TPSB2 (ENSG00000197253), GATA1 (ENSG00000102145), GATA2 (ENSG00000179348) we identified cluster 23 as most likely to be predominantly grouping mast cells (33).

When searching for monocytes clusters we used markers that can identify CD14+ mococytes: CD14 (ENSG00000170458) and CD16+ monocytes: FCGR3A (ENSG00000203747) as well as general monocytes markers: ITGAM (ENSG00000169896) and LYZ (ENSG00000090382) (8). Both had the highest levels of expression in clusters 7 and 9, However we were unable to confidently differentiate between these two types of monocytes. Slight difference in expression levels might suggest that cluster 7 consists mostly of CD16+ monocytes while cluster 9 mostly of CD14+ monocytes as seen on 34.

Furthermore we search for markers that can help us differentiate between cluster 7 and 9 and top two ones were: S100A9 (ENSG00000163220) and S100A8 (ENSG00000143546) which are known monocyte markers and their expression levels were higher in cluster 7. Higher mRNA levels of S100A9 were previously found in CD16+ monocytes (9) (10) and is seen in cluster 7 in our data 35. This brings us to the following identity assignment: cluster 7 - CD16+ monocytes and cluster 9 - CD14+ monocytes.

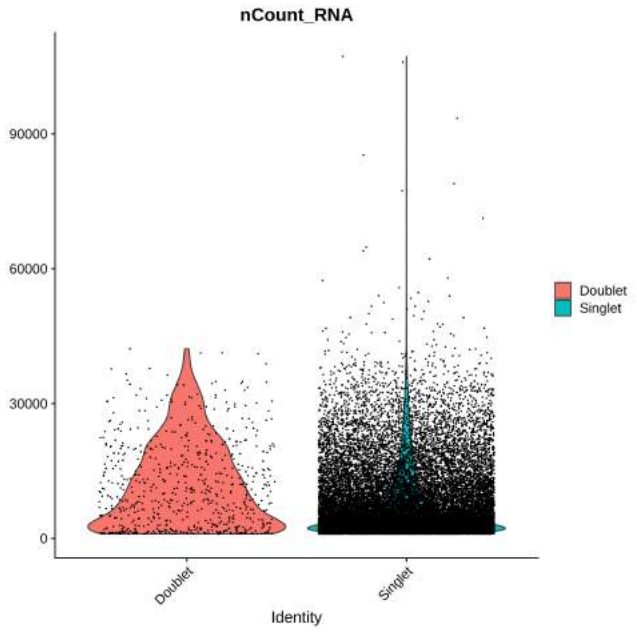


**Fig. 9.** Number of UMI counts in doublets and singlets, $pK = 0.005$

Another set of markers present in the cells are smooth muscle markers such as: ACTG2 (ENSG00000163017), TAGLN (ENSG00000149591), MYH11 (ENSG00000133392), CNN1 (ENSG00000130176), DES (ENSG00000175084), SYNPO2 (ENSG00000172403) 36.

Neuronal cell marker NCAM1 had highest levels of expression in cluster 19 37.

Other clusters that were identified by checking expression level of cell type markers were fibroblasts. To identify fibroblast clusters we use positive markers: THY1 (ENSG00000154096), COL1A1 (ENSG00000108821), COL3A1 (ENSG00000168542), COL1A2 (ENSG00000164692), COL5A1 (ENSG00000130635) (38) and LUM (ENSG00000139329), DCN (ENSG00000011465), LOXL1 (ENSG00000129038), FBLN1 (ENSG00000077942), FBLN2 (ENSG00000163520) (39) and negative marker RGS5 (ENSG00000143248) which is not expressed in fibroblasts (11).
In this case we can observe that clusters identified as fibroblasts can be divided into the ones consisting mostly of normal cells and those consisting mostly of cancer cells, which is visualized on 40 and 41.
To summarize in normal cells most we can find highly expressed fibroblast markers in clusters 8 and 21, while in cancer cell higher markers expression is present in clusters 10 and 12.

***Identity by cluster markers.*** Cluster 1 is characterized by high expression of IgA+ cells markers such as IGHA1 (ENSG00000282633), IGHA2 (ENSG00000276173) and JCHAIN (ENSG00000132465), which can be seen on 58.

Another cluster identity we were able to assign was to cluster 20, which was characterized by high mRNA content of IgG+ cell markers such as IGHG1 (ENSG00000277633), IGHG3 (ENSG00000211897, ENSG00000282184), IGHG4 (ENSG00000277016), IGHGP (ENSG00000253755, ENSG00000282094) 43 . Those markers are most often expressed in IgG+ memory B cells as well as plasma cells (tiller 2007). Other than in cluster 20, high expression levels of IgG+ cell markers were expressed at significant levels in cluster 18, which has been previously identified as B cell cluster. Combining this information suggests that cluster 18 probably contains many different types of B cells.

***Comparison with ScType results.*** Comparing our manual classification of clusters to ScType classification we can see clusters that in both cases had the same identification were 0, 15 as naive CD4+ T cells, 7 as CD16+ monocytes/non-classical monocytes. Cluster 3 was classified as CD8+ T and NK cells, which is consistent with our classification as CD8+ T, NK and $\gamma\delta$ T cells.
In our manual classification we assigned the identity of memory CD4+ T cells to cluster 15, however ScType classified it as naive CD4+ T cells, which is understandable as these cell types are difficult to differentiate and classification depends on choice of markers.
The ScType algorithm classified cluster 1 as naive B cells, which doesn't necessarily exclude presence of IgA+ plasma cells, however it might suggest that rather than plasma cells the cluster 1 population has a higher percentage of different types of IgA+ B cells. Similarly cluster 20 classified by us as IgG+ plasma cells was assigned a naive B cell type by ScType.
The comparison is shown on 44.

Further analysis revealed that some clusters that share cell type identity ( ScType classification) consisted mostly of cells from normal tissue while the other mostly of cells from cancer tissue which is presented 45

***Clustering after correction.*** Again for finding out cluster identity we used top ten most significant markers for each clusters (46).
All clusters that we were able to assign cell identities to can be seen in Table **??**.

***Identity by cell type markers.*** The NK cell markers: NKG7 (ENSG00000105374), GNLY (ENSG00000115523) were highly expressed in cluster 9, 10, 25 and 38 as seen on **??**.
When analyzing $\gamma\delta$ T cells markers: TRDC (ENSG00000211829), TRGC2 (ENSG00000227191) and TRGC1 (ENSG00000211689) we found that clusters 9, 25 and 38 had high expression levels of these markers (48).
Another set of canonicaal markers we analyzed were CD8+ T cell markers: CD8A (ENSG00000153563), CD8B (ENSG00000172116). They were present in the highest levels in 9 and 38 clusters, which can be seen on 49.

In this case CD4+ T cells markers: CD4 (ENSG00000010610), CCR7 (ENSG00000126353), IL7R (ENSG00000168685) allowed us to identify naive CD4+ T (CCR7+) population as mainly cluster 1 and memory CD4+ T (CCR7-) population in clusters 5, 10, 11 and 19. (50).

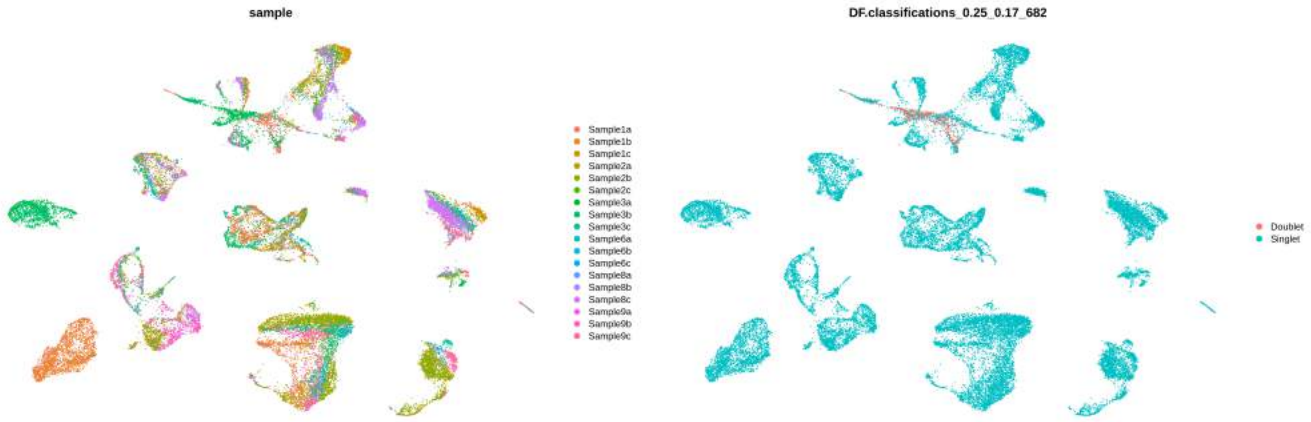


**Fig. 10.** Distribution of doublets between samples, $pK = 0.017$

| Cell type | Clusters |
|---|---|
| NK cells | 9, 10, 25, 38 |
| $\gamma\delta$ T | 9, 25, 38 |
| CD8+ T | 9, 38 |
| Naive CD4+ T | 1 |
| Memory CD4+ T | 5, 10, 11, 19 |
| Monocytes | 5, 11, 19 |
| Mast cells | 33 |
| B cells | 6, 21, 23 |
| Neuronal cells | 22 |
| Fibroblasts | 2, 14, 32, 34 |
| IgA+ plasma cell | 13 |
| Smooth muscle cell | 12, 32, 39 |

**Table 2.** Suggested cluster identity by high expression of cell type markers

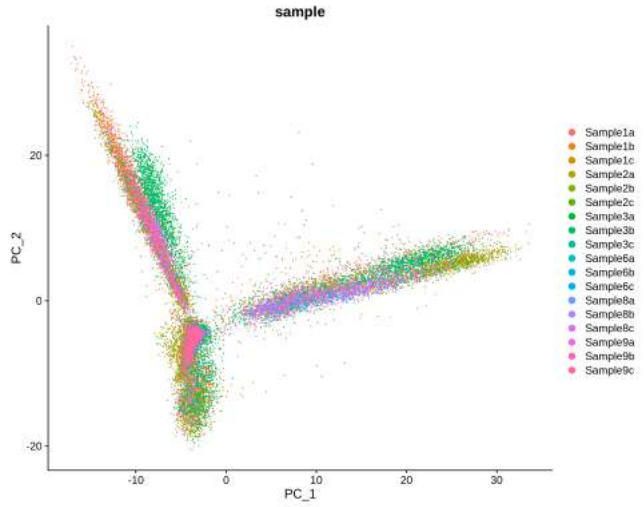


**Fig. 11.** PC1 and PC2 by sample before correction

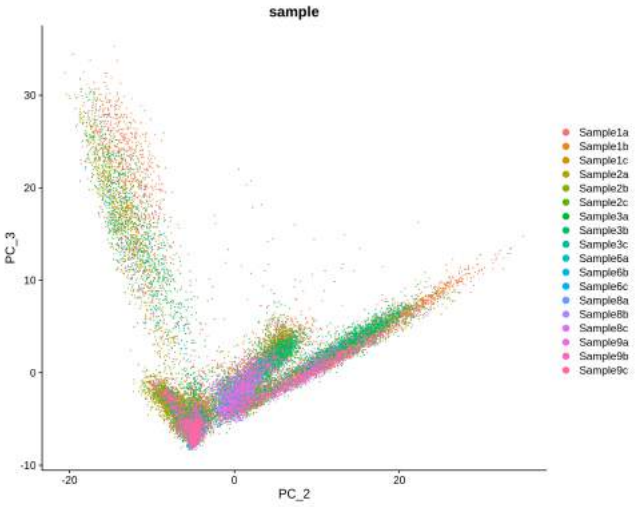


**Fig. 12.** PC2 and PC3 by sample before correction

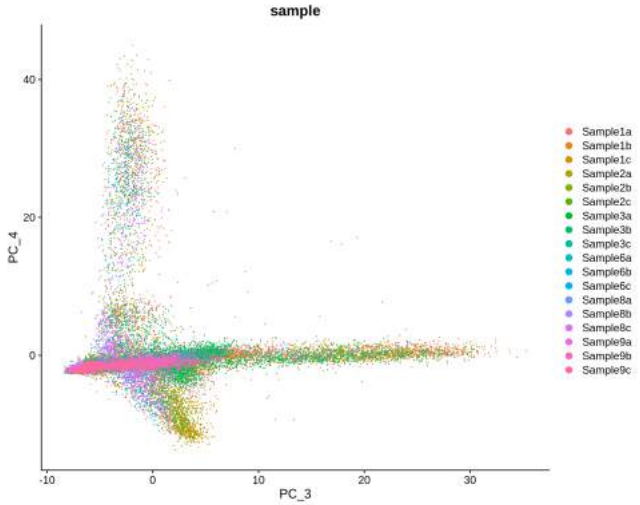


**Fig. 13.** PC3 and PC4 by sample before correction

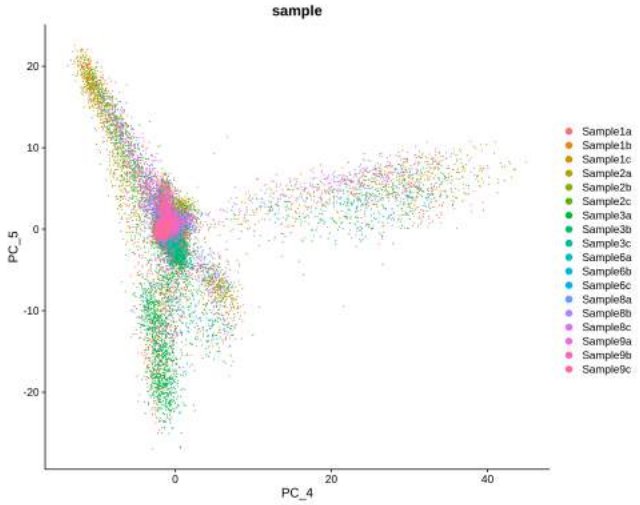


**Fig. 14.** PC4 and PC5 by sample before correction

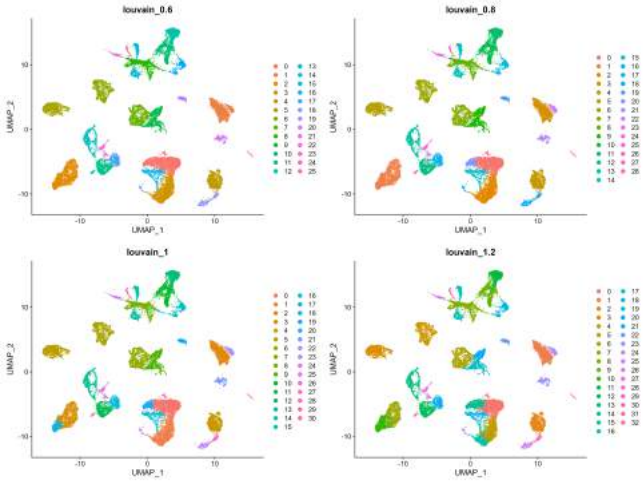


**Fig. 15.** Clustering results depending on resolution before correction

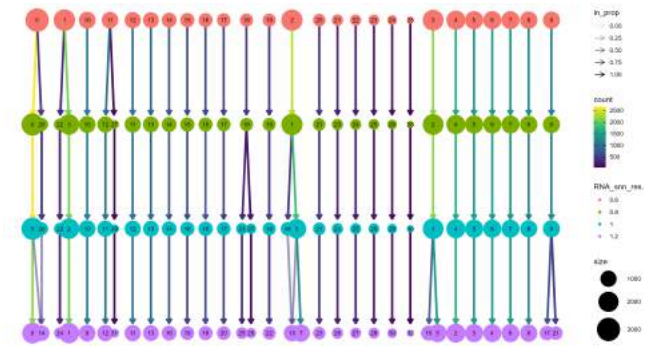

**Fig. 16.** Distribution of cells between clusters depending on resolution before correction

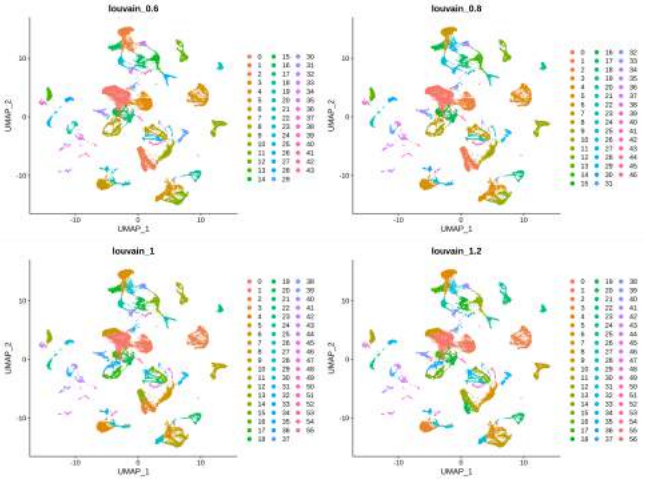


**Fig. 17.** Clustering results depending on resolution after correction

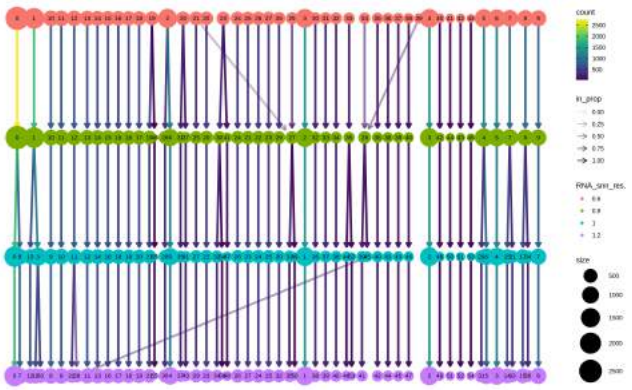

**Fig. 18.** Distribution of cells between clusters depending on resolution after correction

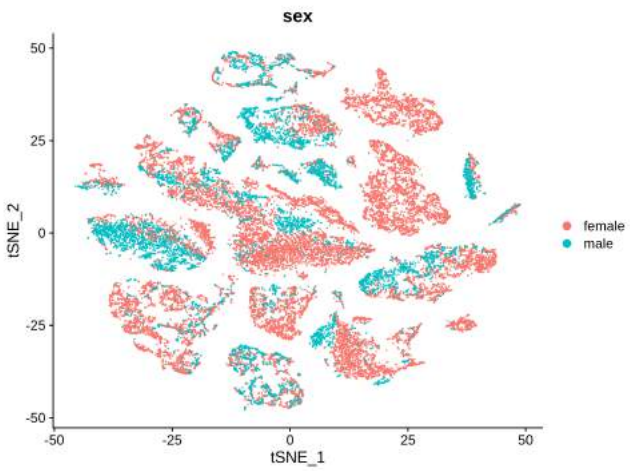


**Fig. 19.** Cell population by sex before correction for sample number

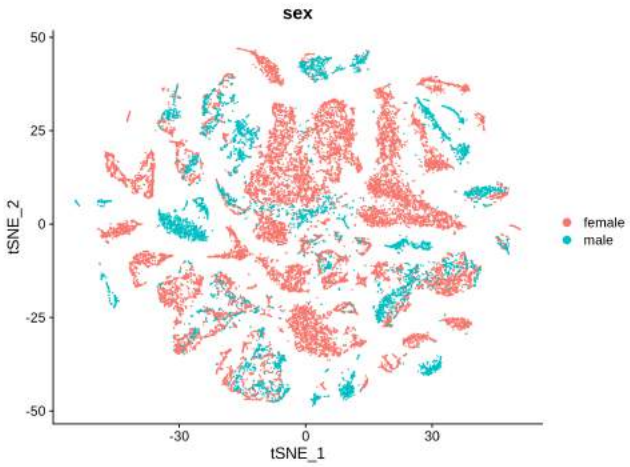


**Fig. 20.** Cell population by sex after correction for sample number

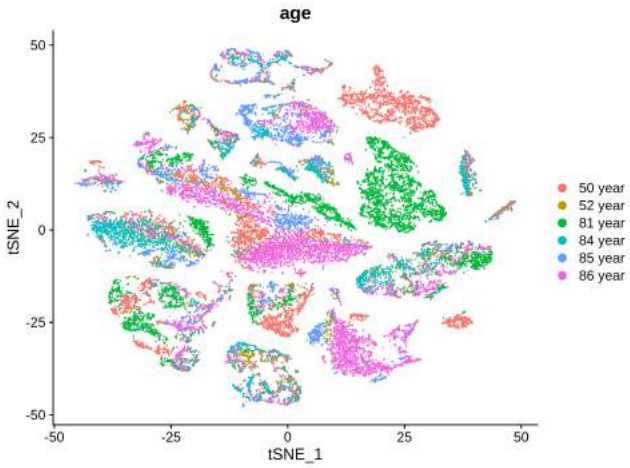


**Fig. 21.** Cell population by age before correction for sample number

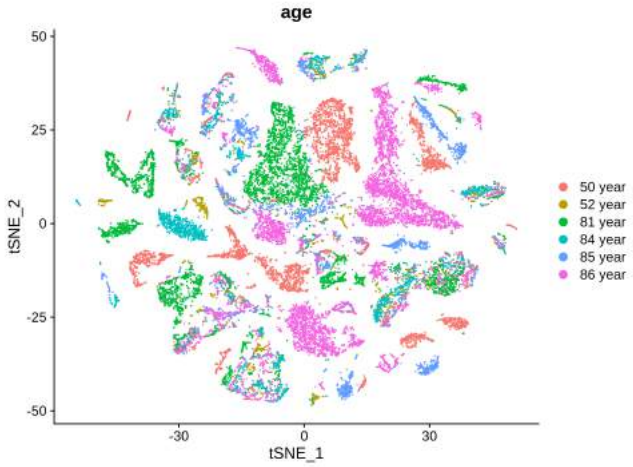


**Fig. 22.** Cell population by age after correction for sample number

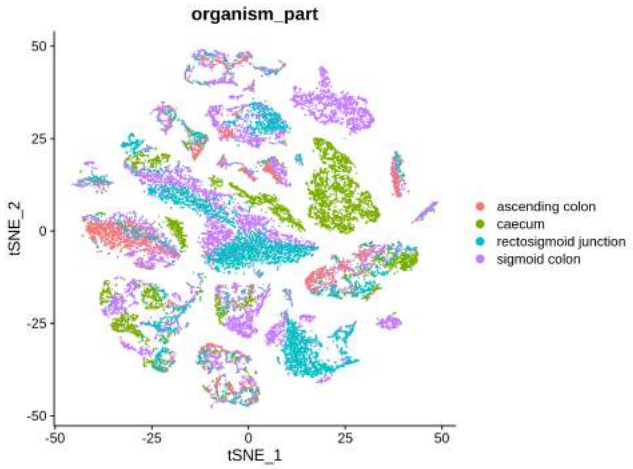


**Fig. 23.** Cell population by organism part from which sample was collected before correction for sample number

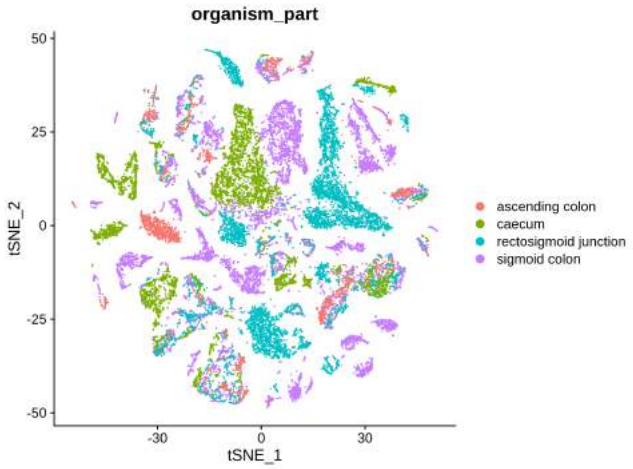


**Fig. 24.** Cell population by organism part from which sample was collected after correction for sample number

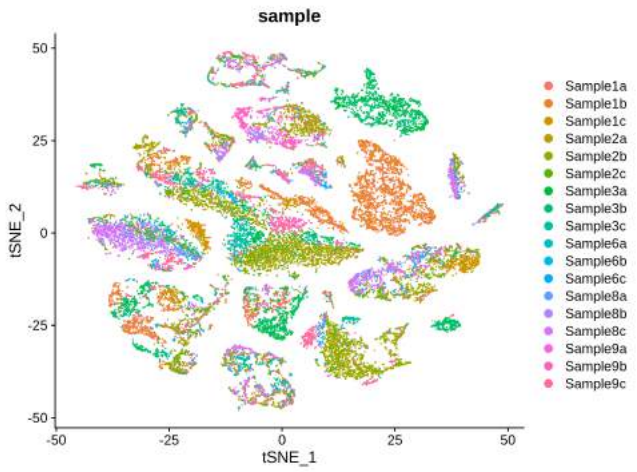


**Fig. 25.** Cell population by sample before correction for sample number

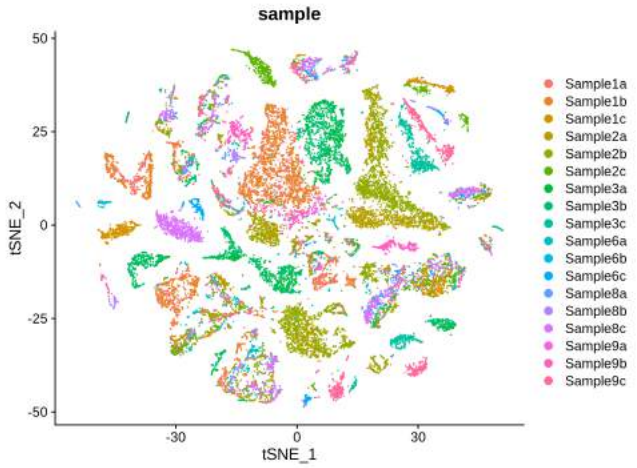


**Fig. 26.** Cell population by sample after correction for sample number

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Cluster 0 | IL7R | TRAC | KLRB1 | CD3D | LTB | LTB | CD2 | TRBC1 | TRBC1 | CD3E |
| Cluster 1 | IGHA1 | IGHA2 | JCHAIN | IGKC | IGLC2 | IGLL5 | MZB1 | IGHA2 | SSR4 | TNFRSF17 |
| Cluster 2 | REG4 | KRT18 | AGR2 | S100P | KRT8 | KRT19 | EPCAM | IFI27 | TSPAN8 | TXN |
| Cluster 3 | NKG7 | GNLY | GZMA | CCL4 | CCL4 | GZMK | CCL5 | CCL5 | CCL4 | GZMB |
| Cluster 4 | IGHM | CD79A | MS4A1 | VPREB3 | CD37 | CD74 | CD79B | HLA-DRA | HLA-DRA | HLA-DPB1 |
| Cluster 5 | MMP7 | DEFB1 | DEFB1 | PSCA | LCN2 | C19orf33 | CEACAM6 | KRT19 | KRT16 | TM4SF4 |
| Cluster 6 | CLDN5 | PLVAP | RAMP2 | VWF | GNG11 | ACKR1 | CCL14 | FABP4 | CAV1 | FABP5 |
| Cluster 7 | SPP1 | S100A9 | CXCL8 | S100A8 | APOC1 | IFI30 | FTL | TYROBP | FCER1G | LYZ |
| Cluster 8 | COL1A1 | COL1A2 | COL3A1 | CTHRC1 | FN1 | BGN | COMP | POSTN | SPARC | TAGLN |
| Cluster 9 | C1QB | C1QA | C1QC | HLA-DPA1 | HLA-DRB1 | HLA-DRA | HLA-DRA | HLA-DPB1 | MS4A6A | HLA-DRB3 |
| Cluster 10 | DPT | MGP | C7 | APOD | GSN | DCN | CCL11 | ADH1B | CCDC80 | FBLN1 |
| Cluster 11 | CD24 | SPINK1 | PRAC1 | FABP1 | GPX2 | TFF3 | RPL12 | RPS6 | RPL31 | RPS18 |
| Cluster 12 | CFD | CFD | IGFBP6 | CLEC3B | DCN | MFAP5 | SFRP2 | MGP | PLAC9 | EFEMP1 |
| Cluster 13 | FABP1 | CA1 | PHGR1 | CKB | MT1G | PIGR | CA2 | LGALS4 | C15orf48 | FXYD3 |
| Cluster 14 | ACTA2 | TAGLN | RGS5 | ADIRF | MYL9 | MUSTN1 | TPM2 | PLN | NDUFA4L2 | CALD1 |
| Cluster 15 | HLA-A | CD3D | TRAC | CD7 | HLA-B | HLA-B | HLA-B | BATF | CD2 | RPS9 |
| Cluster 16 | ADAMDEC1 | CXCL14 | CCL13 | CCL8 | CEBPD | APOE | ABCA8 | CFH | MFAP4 | LTBP4 |
| Cluster 17 | SPINK1 | CEACAM5 | TFF3 | PIGR | LCN2 | FAM3D | CD24 | PLCB4 | ATP1B1 | LEFTY1 |
| Cluster 18 | MS4A1 | TCL1A | CD79B | CD79A | VPREB3 | IRAG2 | RGS13 | MEF2B | IRF8 | PTTG1 |
| Cluster 19 | CRYAB | S100B | GPM6B | CLU | FXYD1 | PLP1 | LGI4 | NRXN1 | ALDH1A1 | PMP22 |
| Cluster 20 | IGHG1 | IGHG3 | IGHG3 | IGHG4 | IGHGP | IGHGP | MZB1 | DERL3 | JSRP1 | DERL3 |
| Cluster 21 | CCN1 | SERPINE1 | IGF1 | CTHRC1 | BGN | C3 | PTGDS | COL5A1 | CCDC80 | SGK1 |
| Cluster 22 | ZG16 | SPINK4 | FCGBP | MUC2 | ITLN1 | MUC2 | FCGBP | WFDC2 | CLCA1 | KLK1 |
| Cluster 23 | TPSB2 | TPSAB1 | CPA3 | CTSG | HPGDS | MS4A2 | HPGD | LTC4S | VWA5A | SLC18A2 |
| Cluster 24 | ACTG2 | DES | CNN1 | MYH11 | MYH11 | PLN | PCP4 | SYNPO2 | LMOD1 | SYNM |
| Cluster 25 | GABBR1 | GABBR1 | GABBR1 | GABBR1 | UBD | UBD | GABBR1 | GABBR1 | GABBR1 | CD2 |

**Fig. 27.** Top 10 significant markers for clusters before correction

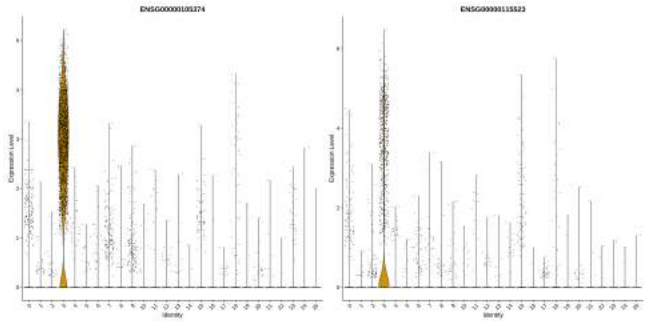

**Fig. 28.** NK cell markers in clusters before correction
NKG7 (ENSG00000105374), GNLY (ENSG00000115523)

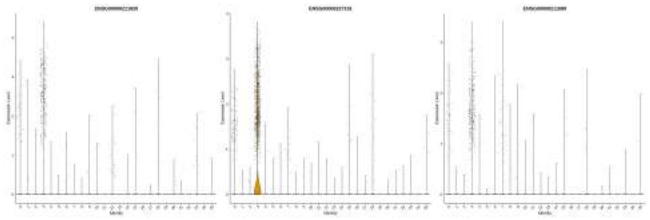

**Fig. 29.** $\gamma\delta$ T cell markers in clusters before correction
TRDC (ENSG00000211829), TRGC2 (ENSG00000227191) and TRGC1 (ENSG00000211689)

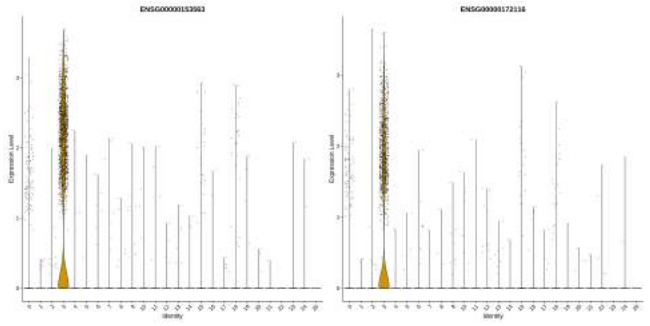

**Fig. 30.** CD8+ T cell markers in clusters before correction
CD8A (ENSG00000153563), CD8B (ENSG00000172116)

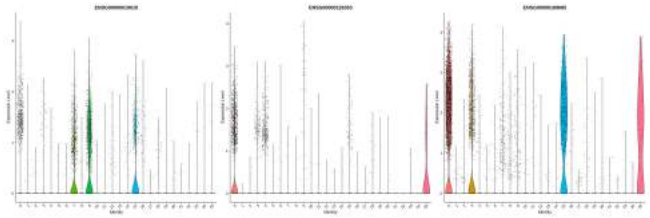

**Fig. 31.** CD4+ T cell markers in clusters before correction
CD4 (ENSG00000010610), CCR7 (ENSG00000126353), IL7R (ENSG00000168685)

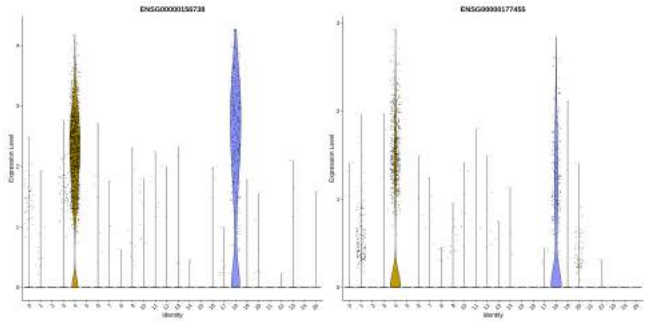

**Fig. 32.** B cell markers in clusters before correction
MS4A1 (ENSG00000156738) and CD19 (ENSG00000177455)

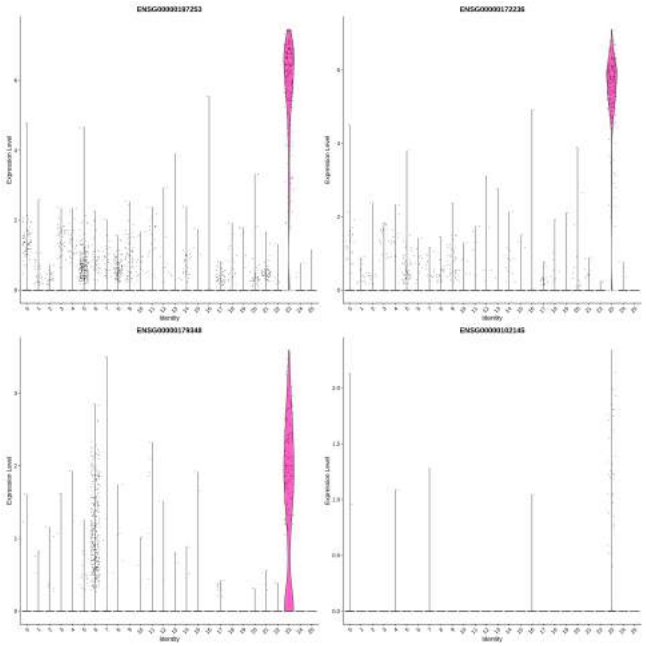

**Fig. 33.** Mast cell markers in clusters before correction
From top left to bottom right: TPSB2 (ENSG00000197253), TPSAB1 (ENSG00000172236), GATA2 (ENSG00000179348), GATA1 (ENSG00000102145)

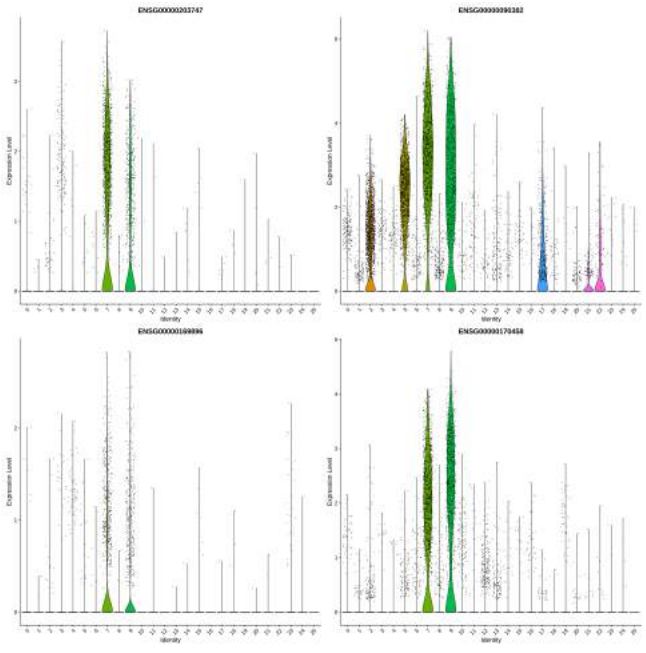

**Fig. 34.** Monocytes markers in clusters before correction
From top left to bottom right: FCGR3A (ENSG00000203747), LYZ (ENSG00000090382), ITGAM (ENSG00000169896), CD14 (ENSG00000170458)

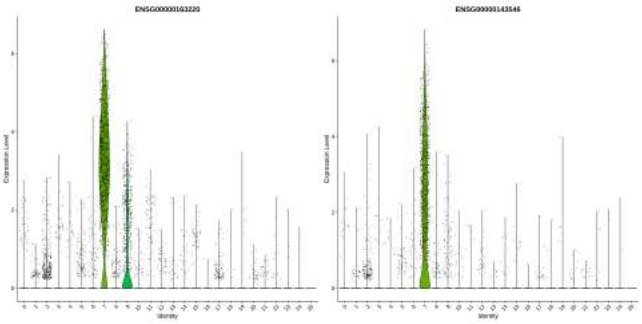

**Fig. 35.** Monocytes markers differentiating 7th and 9th cluster before correction
S100A9 (ENSG00000163220) and S100A8 (ENSG00000143546)

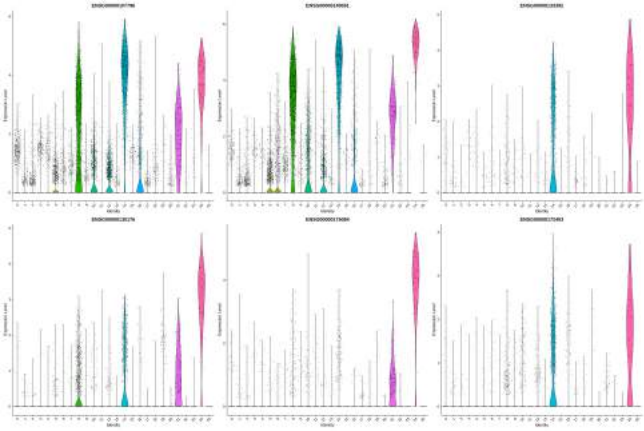

**Fig. 36.** Smooth muscle cells markers in clusters before correction
From top left to bottom right: ACTG2 (ENSG00000163017), TAGLN (ENSG00000149591), MYH11 (ENSG00000133392), CNN1 (ENSG00000130176), DES (ENSG00000175084), SYNPO2 (ENSG00000172403)

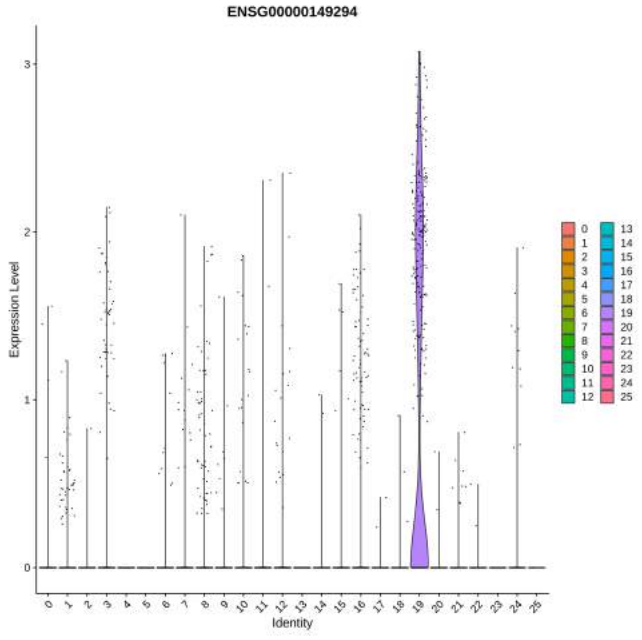


**Fig. 37.** Neuronal cell marker NCAM1 (ENSG00000149294) in clusters before correction

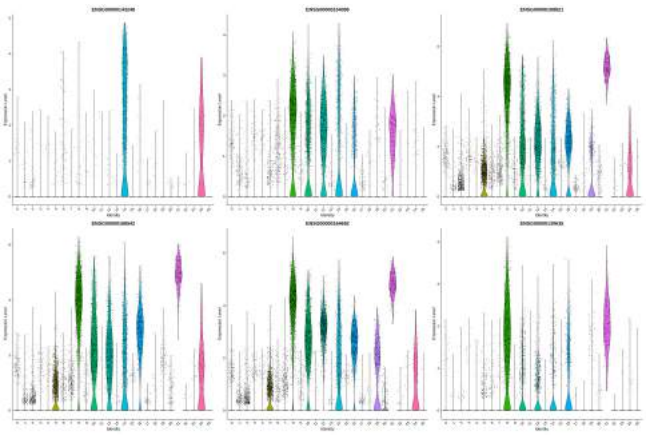

**Fig. 38.** Fibroblasts markers part 1 in clusters before correction
From top left to bottom right: RGS5 (ENSG00000143248), THY1 (ENSG00000154096), COL1A1 (ENSG00000108821), COL3A1 (ENSG00000168542), COL1A2 (ENSG00000164692), COL5A1 (ENSG00000130635)

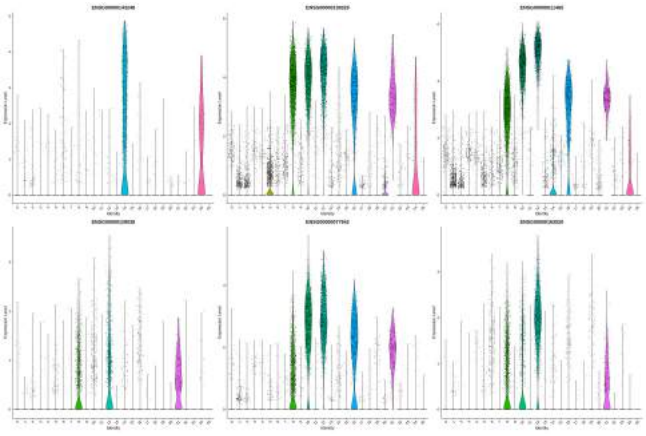

**Fig. 39.** Fibroblasts markers part 2 in clusters before correction
From top left to bottom right: RGS5 (ENSG00000143248), LUM (ENSG00000139329), DCN (ENSG00000011465), LOXL1 (ENSG00000129038), FBLN1 (ENSG00000077942), FBLN2 (ENSG00000163520)

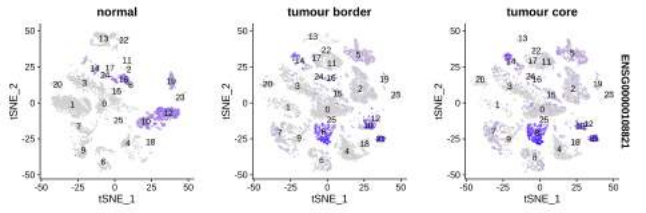


**Fig. 40.** Fibroblasts marker: COL1A1 in clusters grouped by tissue type, before correction

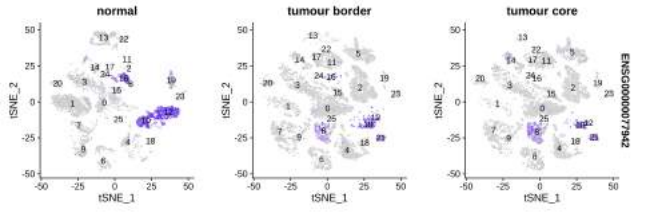


**Fig. 41.** Fibroblasts marker: FBLN1 in clusters grouped by tissue type, before correction

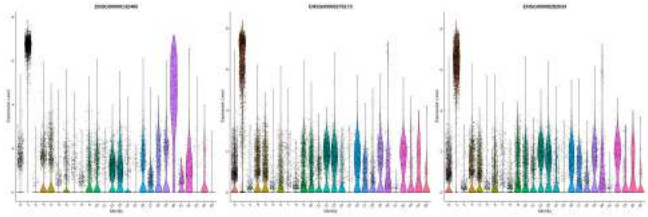

**Fig. 42.** IgA+ plasma cell markers in cluster 1 before correction
JCHAIN (ENSG00000132465), IGHA2 (ENSG00000276173), IGHA1 (ENSG00000282633)

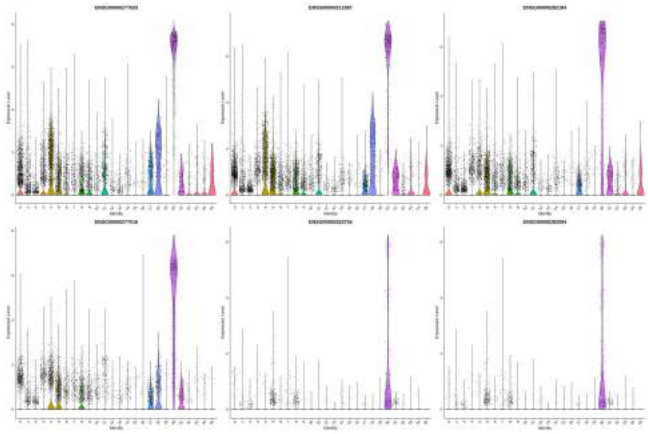

**Fig. 43.** IgG+ plasma cell markers in cluster 20 before correction
IGHG1 (ENSG00000277633), IGHG3 (ENSG00000211897, ENSG00000282184), IGHG4 (ENSG00000277016), IGHGP (ENSG00000253755, ENSG00000282094)

Canonical markers of B cells: MS4A1 (ENSG00000156738), CD19 (ENSG00000177455) were found in clusters 6, 21 and 23 as seen on 51.

After correction for sample number mast cell markers: TPSAB1 (ENSG00000172236), TPSB2 (ENSG00000197253), GATA1 (ENSG00000102145), GATA2 (ENSG00000179348) had highest expression levels in cluster 33 as seen on 52.

While checking in which clusters high levels of monocytes marker mRNA were present we found it hard to differentiate between CD14+ and CD16+ clusters. Markers used were then same as before CD14+ mococytes: CD14 (ENSG00000170458), CD16+ monocytes: FCGR3A (ENSG00000203747) general monocytes markers: ITGAM (ENSG00000169896) and LYZ (ENSG00000090382).
The highest levels of marker mRNA were present in clusters 5, 11 and 19 as can be seen on 53.
For comparison to data before correction we checked expression levels of smooth muscle markers: ACTG2 (ENSG00000163017), TAGLN (ENSG00000149591), MYH11 (ENSG00000133392), CNN1 (ENSG00000130176), DES (ENSG00000175084), SYNPO2 (ENSG00000172403) and found they were present most consistently in clusters 12, 32, 39 (54).

Neuronal cell marker NCAM1 was found highly expressed in cluster 22 55.

Another cell type we checked for comparison with clustering results before correction for sample were fibroblasts. Again we used positive markers: THY1 (ENSG00000154096),

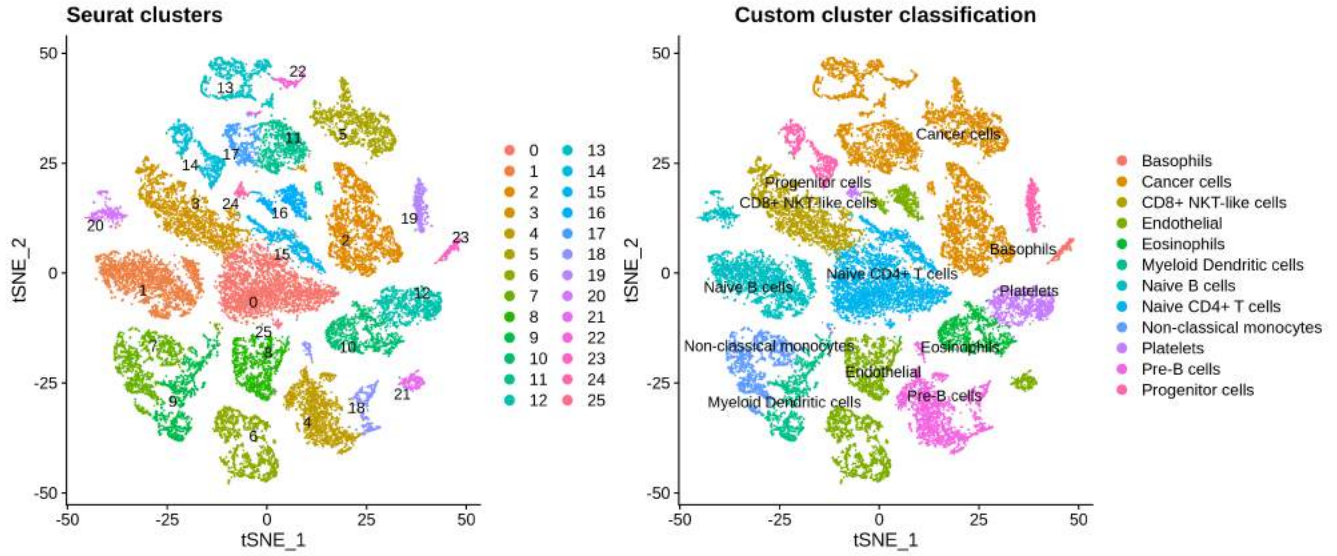


**Fig. 44.** Seurat clusters vs ScType cluster identities before correction

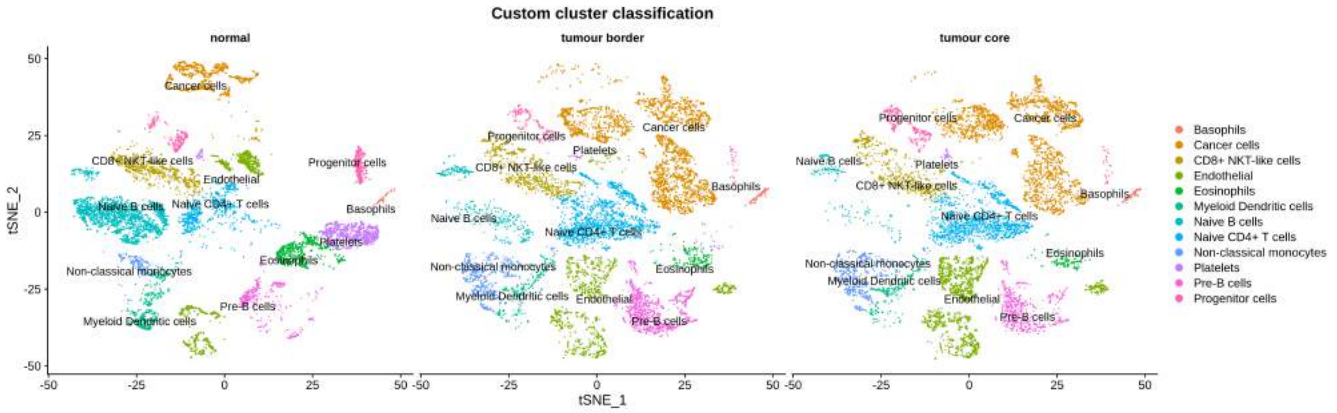


**Fig. 45.** ScType cluster identities by sampling site before correction

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Cluster 0 | RGS4 | AQR2 | KRT18 | S100P | KRT8 | EPCAM | KRT19 | IFI27 | TSPAN8 | TXN |
| Cluster 1 | TRAC | LTB | LTB | IL7R | BIRC3 | CD3D | TRBC1 | TNFRSF4 | BTG1 | TRBC1 |
| Cluster 2 | CFD | CFD | DCN | SFRP2 | IGFBP6 | MGP | CLEC3B | MFAP5 | FBLN1 | GSN |
| Cluster 3 | MMP7 | DEFB1 | DEFB1 | PSCA | LCN2 | CEACAM6 | C19orf33 | KRT19 | KRT18 | TM4SF4 |
| Cluster 4 | CLDN5 | PLVAP | RAMP2 | VWF | GNG11 | ACKR1 | CCL14 | FABP4 | CAV1 | PECAM1 |
| Cluster 5 | C1QB | C1QA | HLA-DPA1 | C1QC | HLA-DRB1 | LYZ | TYROBP | HLA-DRA | HLA-DRA | MS4A6A |
| Cluster 6 | IGHM | CD79A | MS4A1 | CD37 | HLA-DPB1 | VPREB3 | CD74 | HLA-DRA | HLA-DRA | CD79B |
| Cluster 7 | KLRB1 | CD69 | CD7 | CD3D | IL7R | CD52 | EVL | IL32 | TRAC | CD3 |
| Cluster 8 | FABP1 | CA1 | PHGR1 | CK5 | MT1G | PIGR | CA2 | LGALS4 | C12orf45 | FXYD3 |
| Cluster 9 | GZMK | NKG7 | CCL4 | GZMA | DUSP2 | CST7 | CCL5 | CCL5 | CCL4 | CCL4 |
| Cluster 10 | HLA-A | CD7 | HLA-B | HLA-B | HLA-B | CD3D | TRAC | CD2 | BAT1 | GZMA |
| Cluster 11 | CXCL8 | S100A9 | SPP1 | IL1B | S100A8 | FCER1G | IFI30 | FTL | HLA-DRA | TYROBP |
| Cluster 12 | ACTA2 | TAGLN | RGS5 | ADIRF | MYL9 | MUSTN1 | TPM2 | PLN | NDUFA4L2 | IGFBP7 |
| Cluster 13 | JCHAIN | IGHA1 | IGHA2 | IGHA2 | MZB1 | SSR4 | TNFRSF17 | ANKRD36 | UBE2J1 | ANKRD36BP2 |
| Cluster 14 | COL1A1 | FN1 | COMP | COL3A1 | COL1A2 | CTHRC1 | TAGLN | SFRP2 | POSTN | AEBP1 |
| Cluster 15 | CD24 | FABP1 | PCP4 | RPS17 | PRAC1 | RPL12 | RPS21 | TNNC2 | RPS18 | SPINK1 |
| Cluster 16 | DPT | C7 | CCL11 | MGP | GSN | THBS4 | ADH1B | DCN | LUM | CXCL12 |
| Cluster 17 | CST1 | CTHRC1 | COL1A1 | BGN | SFRP4 | COL1A2 | IGFBP5 | COL3A1 | POSTN | LUM |
| Cluster 18 | SPINK1 | CEACAM5 | TFF3 | LCN2 | PIGR | RAMP3 | CD24 | LEFTY1 | PLCB4 | PRAC1 |
| Cluster 19 | APOC1 | LYZ | S100A9 | S100A8 | TYROBP | HLA-DRB1 | FCER1G | IFI30 | ENSG00000268173 | AIF1 |
| Cluster 20 | JCHAIN | IGHA2 | MZB1 | IGHA1 | TNFRSF17 | DERL3 | DERL3 | CCR10 | GNG7 | NR3I |
| Cluster 21 | MS4A1 | TCL1A | CD79B | VPREB3 | IRAG2 | RGS13 | MEF2B | IRF8 | PTTG1 | HLA-DQA1 |
| Cluster 22 | CRYAB | S100B | GPM6B | CLU | FXYD1 | PLP1 | LGI4 | NRGN1 | ALDH1A1 | PMP22 |
| Cluster 23 | IGHM | VPREB3 | MS4A1 | CD79A | CD79B | BANK1 | CD19 | HLA-DQB1 | HLA-DQB1 | NAPSB |
| Cluster 24 | ADAMDEC1 | CCL13 | CXCL14 | CCL8 | CCL11 | APOE | A2CA6 | CFH | MFAP4 | LTBP4 |
| Cluster 25 | KLRB1 | GZMA | NKG7 | CD3D | CD7 | CD2 | CD3E | GZMM | CST7 | GZMH |
| Cluster 26 | JCHAIN | MZB1 | IGHA2 | DERL3 | CD79A | PIM2 | TNFRSF17 | DERL3 | TXNDC5 | TENT5C |
| Cluster 27 | HLA-A | HLA-C | HLA-B | HLA-B | CD3D | HLA-C | B2M | CTLA4 | CD2 | TRAC |
| Cluster 28 | MZB1 | IGHA2 | IGHA1 | TNFRSF17 | DERL3 | DERL3 | ANKRD36BP2 | TXNDC5 | TENT5C | GNG7 |
| Cluster 29 | HLA-A | IL17F | IL17A | CD2 | CD3D | IL26 | CD3E | TRBC2 | CXCR6 | KLRB1 |
| Cluster 30 | GABBR1 | GABBR1 | GABBR1 | GABBR1 | UBD | UBD | UBD | UBD | UBD | GABBR1 |
| Cluster 31 | ZG16 | SPINK4 | FCGBP | MUC2 | ITLN1 | MUC2 | FCGBP | WFDC2 | CLGA1 | KLK1 |
| Cluster 32 | CCN1 | SERPINB1 | IGF1 | CTHRC1 | BGN | C3 | PTGDS | CCDC80 | COL5A1 | SOX1 |
| Cluster 33 | TPSB2 | TPSAB1 | CPA3 | CTSG | HPGDS | MS4A2 | HPGD | LTC4S | VWA5A | SLC18A2 |
| Cluster 34 | CXCL14 | F3 | PDGFRA | FRZB | INSC1 | EDNRB | HSD17B2 | PPP1R14A | IBSP | SOX6 |
| Cluster 35 | MZB1 | DERL3 | IGLV3-1 | TNFRSF17 | FCRL5 | DERL3 | CIBAR2 | SPAG4 | ANKRD36BP2 | JSRP1 |
| Cluster 36 | DERL3 | MZB1 | FCRL5 | CD38 | DERL3 | IGLV3-1 | SPAG4 | TNFRSF17 | IGHG4 | IGHGP |
| Cluster 37 | TNFRSF17 | HLA-DQA1 | IGHA2 | SPINK2 | MZB1 | DERL3 | DERL3 | ABCB9 | TXNDC5 | ANKRD28 |
| Cluster 38 | TRGC2 | ENSG00000255691 | KLRC3 | KLRC1 | RPS4Y1 | KLRC4-KLRK1 | KLRC2 | KLRK1 | CD69 | CCL5 |
| Cluster 39 | ACTG2 | DES | CNN1 | MYH11 | MYH11 | PLN | PCP4 | SYNPO2 | LMOD1 | SYNM |
| Cluster 40 | DERL3 | TNFRSF17 | DERL3 | MZB1 | ANKRD36BP2 | IGHA2 | TXNDC5 | BLOC1S5-TXNDC5 | JCHAIN | FKBP11 |
| Cluster 41 | IGHGP | IGHGP | CIBAR2 | JSRP1 | IGLV1-51 | IGLV6-57 | FCRL3 | DERL3 | MZB1 | IGLV3-1 |
| Cluster 42 | CXCL8 | RPS27 | IL7R | KLRB1 | H3F3B | RPL32 | RPL10 | RPS15A | RPL39 | EEF1D |
| Cluster 43 | DERL3 | DERL3 | MZB1 | TNFRSF17 | FCRL5 | ANKRD36BP2 | TXLNGY | IGHA2 | H2BC8 | CD79A |

**Fig. 46.** Top 10 significant markers for clusters after correction

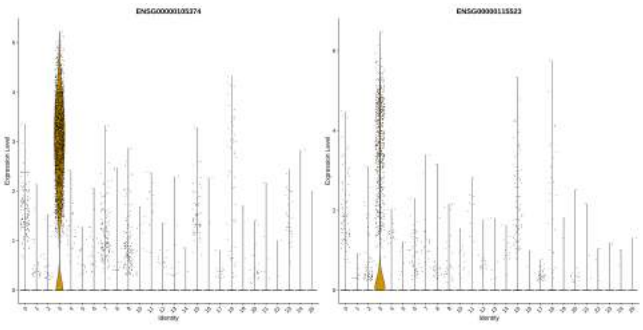

**Fig. 47.** NK cell markers in clusters before correction
NKG7 (ENSG00000105374), GNLY (ENSG00000115523)



**Fig. 48.** $\gamma\delta$ T cell markers in clusters after correction
TRDC (ENSG00000211829), TRGC2 (ENSG00000227191) and TRGC1 (ENSG00000211689)

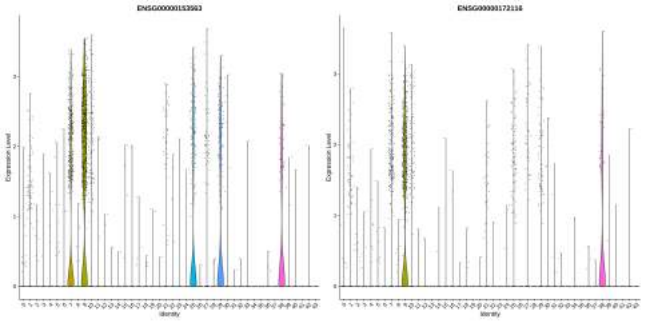

**Fig. 49.** CD8+ T cell markers in clusters after correction
CD8A (ENSG00000153563), CD8B (ENSG00000172116)

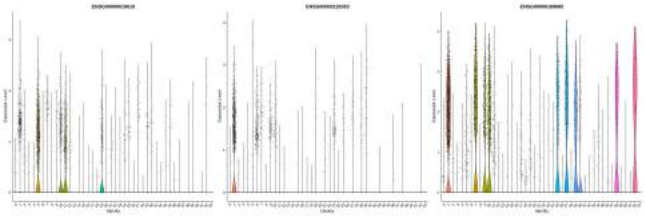

**Fig. 50.** CD4+ T cell markers in clusters after correction
CD4 (ENSG00000010610), CCR7 (ENSG00000126353), IL7R (ENSG00000168685

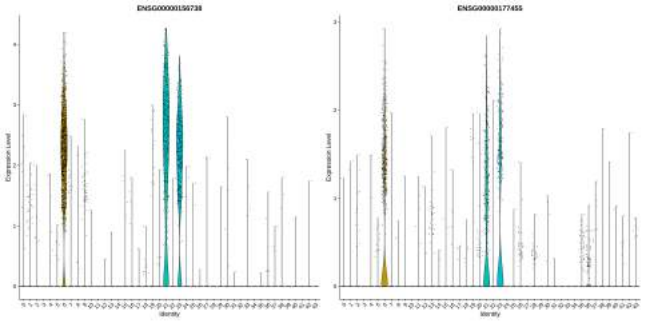

**Fig. 51.** B cell markers in clusters after correction
MS4A1 (ENSG00000156738) and CD19 (ENSG00000177455)

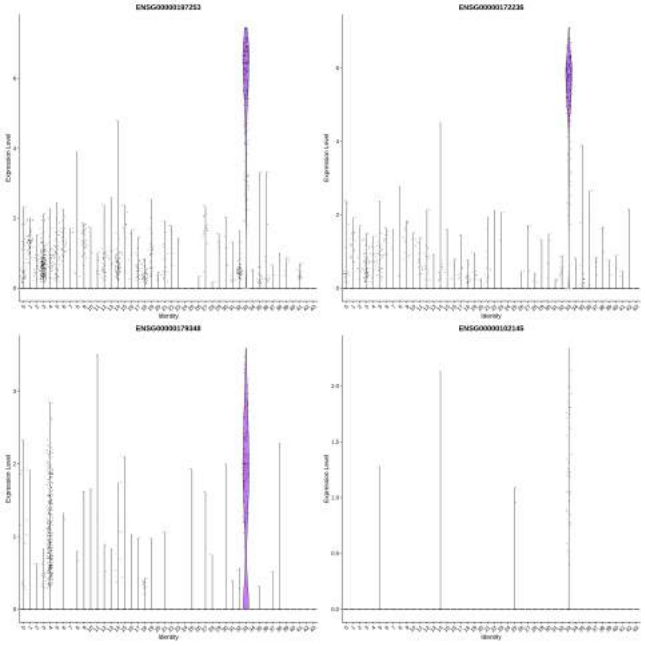

**Fig. 52.** Mast cell markers in clusters after correction
From top left to bottom right: TPSB2 (ENSG00000197253), TPSAB1 (ENSG00000172236), GATA2 (ENSG00000179348), GATA1 (ENSG00000102145)

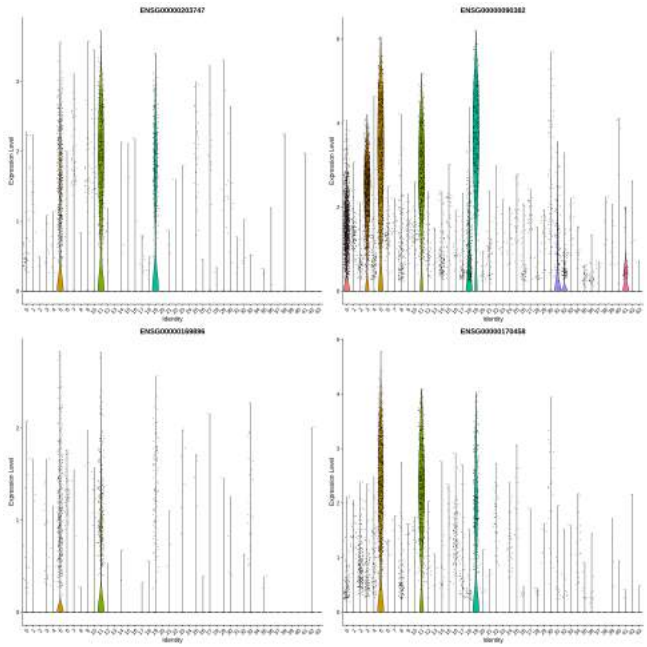

**Fig. 53.** Monocytes markers in clusters after correction
From top left to bottom right: FCGR3A (ENSG00000203747), LYZ (ENSG00000090382), ITGAM (ENSG00000169896), CD14 (ENSG00000170458)

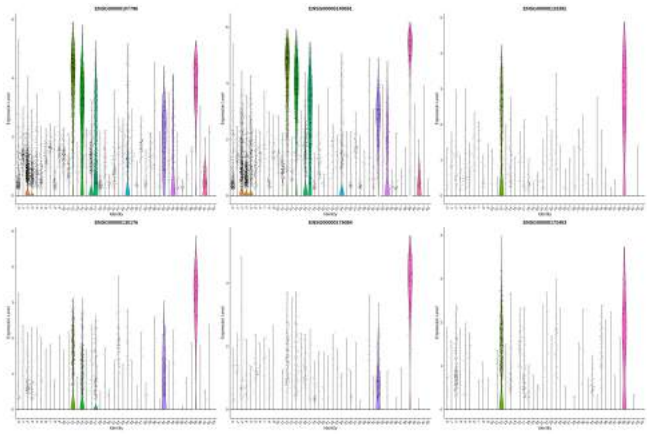

**Fig. 54.** Smooth muscle cells markers in clusters before correction
From top left to bottom right: ACTG2 (ENSG00000163017), TAGLN (ENSG00000149591), MYH11 (ENSG00000133392), CNN1 (ENSG00000130176), DES (ENSG00000175084), SYNPO2 (ENSG00000172403)

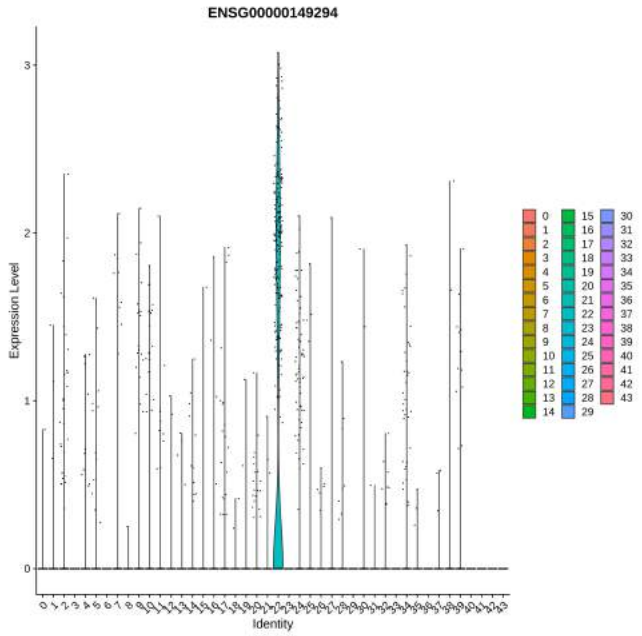


**Fig. 55.** Neuronal cell marker NCAM1 (ENSG00000149294) in clusters aftercorrection

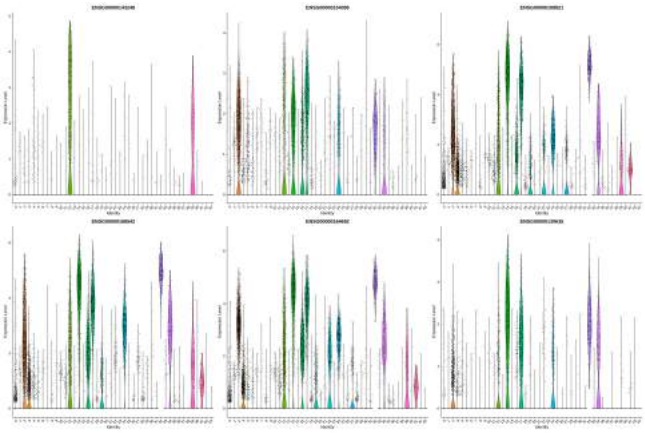

**Fig. 56.** Fibroblasts markers part 1 in clusters after correction
From top left to bottom right: RGS5 (ENSG00000143248), THY1 (ENSG00000154096), COL1A1 (ENSG00000108821), COL3A1 (ENSG00000168542), COL1A2 (ENSG00000164692), COL5A1 (ENSG00000130635)

COL1A1 (ENSG00000108821), COL3A1 (ENSG00000168542), COL1A2 (ENSG00000164692), COL5A1 (ENSG00000130635) (56) and LUM (ENSG00000139329), DCN (ENSG00000011465), LOXL1 (ENSG00000129038), FBLN1 (ENSG00000077942), FBLN2 (ENSG00000163520) (57) and negative marker RGS5 (ENSG00000143248) which is not expressed in fibroblasts (11). Markers expression indicates that fibroblast population is most likely located in clusters 2, 14, 32 and 34.

***Identity by cluster markers.*** Similarly as in clustering results before data correction we were able to identify IgA+ plasma cells by analyzing markers IGHA1 (ENSG00000282633), IGHA2 (ENSG00000276173) and JCHAIN (ENSG00000132465) as markers which were identifying cluster 13.

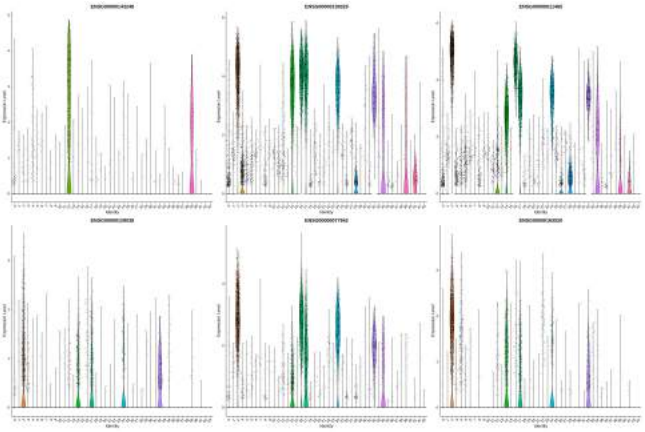

**Fig. 57.** Fibroblasts markers part 2 in clusters after correction
From top left to bottom right: RGS5 (ENSG00000143248), LUM (ENSG00000139329), DCN (ENSG00000011465), LOXL1 (ENSG00000129038), FBLN1 (ENSG00000077942), FBLN2 (ENSG00000163520)

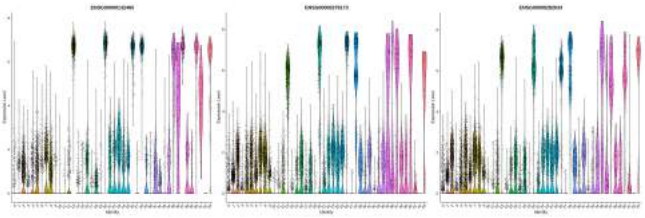

**Fig. 58.** IgA+ plasma cell markers in cluster 13 before correction
JCHAIN (ENSG00000132465), IGHA2 (ENSG00000276173), IGHA1 (ENSG00000282633)

***Comparison with ScType results.*** In this case the classification diverged. However most classification were similar enough that the differences can be explained by choice of markers.
In both cases cluster 1 was assigned naive CD4+ T cell type. Clusters that were classified by us as B cells, ScType classified as Pre-B cells and as naive B cells clusters 43, 40, 28, 13, 37 and 35. So our classification of cluster 13 as IgA+ plasma cells was most likely correct.
As non-classical monocytes ScType marked only clyster 19, while cluster 11 was identified as macrophages and cluster 5 as dendric cells.
Further comparison can be seen on 59 and the distribution of clusters between tissues is shown on 60.

## Discussion

Single-cell ‘RNA sequencing analysis is a complex subject that without depending on prior knowledge about the cells and their expression patterns as well as markers that might help to identify them. For that reason most research tries to merge information from different types of sequencing for example bulk RNA sequencing and sc-RNA-seq. In cancer research this approach allows to classify cells by their CMS type. However, if one studies only sc-RNA-seq data, classification of unsupervised clusters is a multifaceted problem that requires choosing appropriate markers to differentiate groups based on current knowledge about transcriptomics map of cancer tissues. This clearly underlines the need of

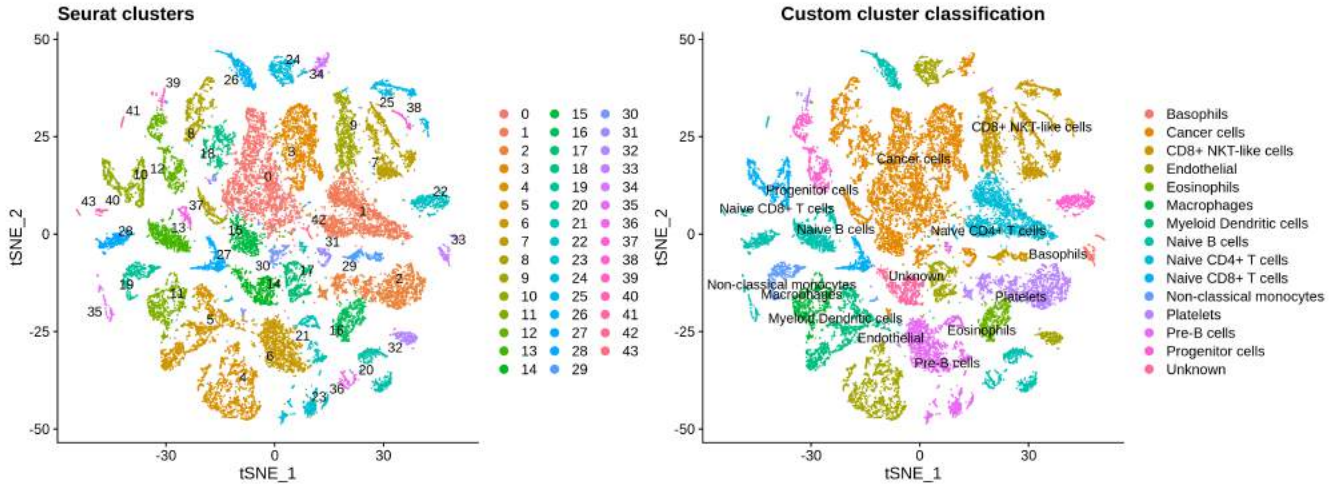


**Fig. 59.** Seurat clusters vs ScType cluster identities after correction

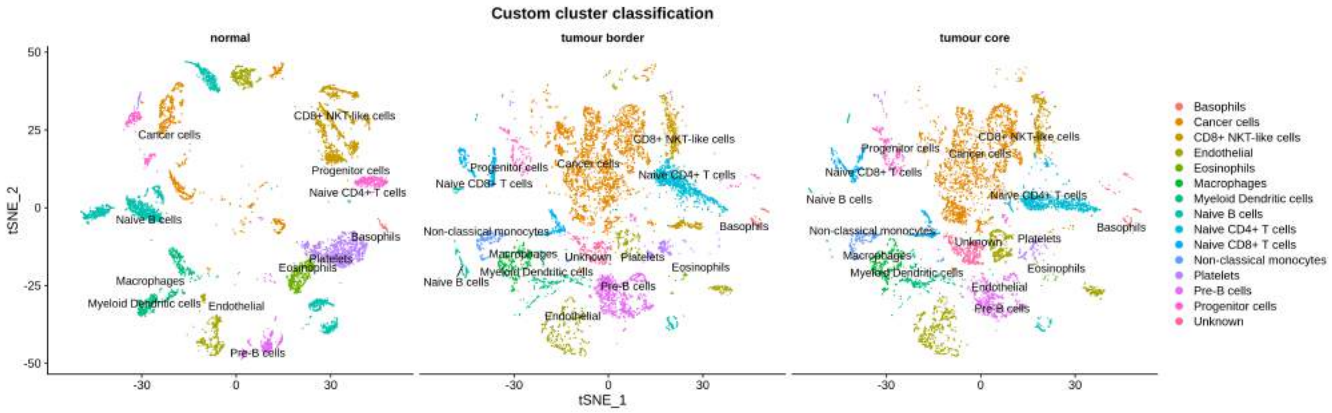


**Fig. 60.** ScType cluster identities by sampling site after correction

such databases as Protein Atlas.

Generally identification of cell populations relies on manual annotation of cell clusters using established transcriptomic markers. It is unquestionably a time consuming process and relies on proper selection of markers for specific dataset. Finding informative markers for both individual cell clusters as well as various cell types present in the sample.

Using platforms such as ScType that enable fully automated cell type classification is definitely a more effective approach, especially if we are interested in general classification that would later allow us to explore cell subtypes further.

Another important thing that can be noted based on our results is the fact that as we previously stated, trying to correct data for possible batch effects/variation from non biological variables is not always beneficial for the sake of interpretability of our results. After regressing our data for sample number we obtained a larger number of simmingly better separated clusters, however it didn't help with the classification. We can even say that assigning cluster identity was easier before data correction.

All of the above observations only confirm that sc-RNA-seq data analysis and unsupervised clustering are not straightforward problems and require a lot of thought and background knowledge about the data and studied processes.

## References


1. Hae-Ock Lee, Yourae Hong, Hakki Emre Etlioglu, Yong Beom Cho, Valentina Pomella, Ben Van den Bosch, Jasper Vanhecke, Sara Verbandt, Hyekyung Hong, Jae-Woong Min, Nayoung Kim, Hye Hyeon Eum, Junbin Qian, Bram Boeckx, Diether Lambrechts, Petros Tsantoulis, Gert De Hertogh, Woosung Chung, Taeseob Lee, Minae An, Hyun-Tae Shin, Je-Gun Joung, Min-Hyeok Jung, Gunhwan Ko, Pratyaksha Wirapati, Seok Hyung Kim, Hee Cheol Kim, Seong Hyeon Yun, Iain Bee Huat Tan, Bobby Ranjan, Woo Yong Lee, Tae-You Kim, Jung Kyoon Choi, Young-Joon Kim, Shyam Prabhakar, Sabine Tejpar, and Woong-Yang Park. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nature Genetics*, 52(6):594–603, June 2020. ISSN 1546-1718. doi: 10.1038/s41588-020-0636-z.
2. Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019. ISSN 1744-4292. doi: 10.15252/msb.20188746.
3. Xiaoyan Qiu, Xiaohui Zhu, Liang Zhang, Yuntao Mao, Jian Zhang, Peng Hao, Guohui Li, Peng Lv, Zhixin Li, Xin Sun, Lemeng Wu, Jie Zheng, Yuqing Deng, Chunmei Hou, Peixian Tang, Shuren Zhang, and Youhui Zhang. Human Epithelial Cancers Secrete Immunoglobulin G with Unidentified Specificity to Promote Growth and Survival of Tumor Cells1. *Cancer Research*, 63(19):6488–6495, October 2003. ISSN 0008-5472.
4. Thiago T. Maciel, Ivan C. Moura, and Olivier Hermine. The role of mast cells in cancers. *F1000Prime Reports*, 7:09, January 2015. ISSN 2051-7599. doi: 10.12703/P7-09.
5. Hui Meng, Wencai Li, Lisa A. Boardman, and Liang Wang. Loss of ZG16 is associated with molecular and clinicopathological phenotypes of colorectal cancer. *BMC Cancer*, 18:433, April 2018. ISSN 1471-2407. doi: 10.1186/s12885-018-4337-2.
6. Aleksandr Ianevski, Anil K. Giri, and Tero Aittokallio. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature Communications*, 13(1):1246, March 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28803-w. Number: 1 Publisher: Nature Publishing Group.
7. Max Karlsson, Cheng Zhang, Loren Méar, Wen Zhong, Andreas Digre, Borbala Katona, Evelina Sjöstedt, Lynn Butler, Jacob Odeberg, Philip Dusart, Fredrik Edfors, Per Oksvold, Kalle von Feilitzen, Martin Zwahlen, Muhammad Arif, Ozlem Altay, Xiangyu Li, Mehmet Ozcan, Adil Mardonoglu, Linn Fagerberg, Jan Mulder, Yonglun Luo, Fredrik Ponten, Mathias Uhlén, and Cecilia Lindskog. A single–cell type transcriptomics map of human tissues. *Science Advances*, 7(31):eabh2169, July 2021. ISSN 2375-2548. doi: 10.1126/sciadv.abh2169.

8. Theodore S. Kapellos, Lorenzo Bonaguro, Ioanna Gemünd, Nico Reusch, Adem Saglam, Emily R. Hinkley, and Joachim L. Schultze. Human Monocyte Subsets and Phenotypes in Major Chronic Inflammatory Diseases. *Frontiers in Immunology*, 10, 2019. ISSN 1664-3224.
9. Xiaohe Yan, Pål Andresen, Xhevat Lumi, Qingshan Chen, and Goran Petrovski. Expression of Progenitor Cell Markers in the Glial-Like Cells of Epiretinal Membranes of Different Origins. *Journal of Ophthalmology*, 2018:e7096326, December 2018. ISSN 2090-004X. doi: 10.1155/2018/7096326.
10. Fernando O Martinez. The transcriptome of human monocyte subsets begins to emerge. *Journal of Biology*, 8(11):99, 2009. ISSN 1478-5854. doi: 10.1186/jbiol206.
11. Lars Muhl, Guillem Genové, Stefanos Leptidis, Jianping Liu, Liqun He, Giuseppe Mocci, Ying Sun, Sonja Gustafsson, Byambajav Buyandelger, Indira V. Chivukula, Åsa Segerstolpe, Elisabeth Raschperger, Emil M. Hansson, Johan L. M. Björkegren, Xiao-Rong Peng, Michael Vanlandewijck, Urban Lendahl, and Christer Betsholtz. Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. *Nature Communications*, 11(1):3953, August 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17740-1.