**3.4 Annotating clusters**

Cell clustering allows unsupervised definition of cell types.

We need to be able to annotate the identified cell clusters.

This is non-trivial and often requires an expert in the tissue being studied.

Computational tools can leverage large volumes of existing single-cell data to accelerate the process of cell type annotation.

**3.4.1 Different approaches to cell type annotation**

Benchmarking study Abdelaal et al. 2019 <https://doi.org/10.1186/s13059-019-1795-z>

Review Pasquini et al. 2021 <https://doi.org/10.1016/j.csbj.2021.01.015>

One can annotate the cell types of individual cells or of clusters of cells.

Briefly, there are three main computational approaches to annotate cell types (see Figure below)

1. **Marker-gene-database-based** approaches use sets of marker genes from the literature and previous single-cell studies. The sets of marker genes can distinguish different cell types. Single cells or clusters are scored using these marker gene sets estimate the overall expression levels of these genes. See the Seurat AddModuleScore function (<https://www.waltermuskovic.com/2021/04/15/seurat-s-addmodulescore-function/>) Some heuristics and scoring criteria are then applied to assign the cell type.
2. **Correlation-based** approaches apply multiple correlation measures to estimate the similarity between the single cells / clusters in the input dataset against some reference data e.g. single-cell atlases or bulk RNA-seq databases such as GTeX / FANTOM.
3. **Supervised-classification-based** approaches use supervised learning to predict cell type labels. The learning model is first trained on some single-cell reference atlas before being applied onto the input dataset to compute the probability that a single cell / cluster belongs to a particular cell type.

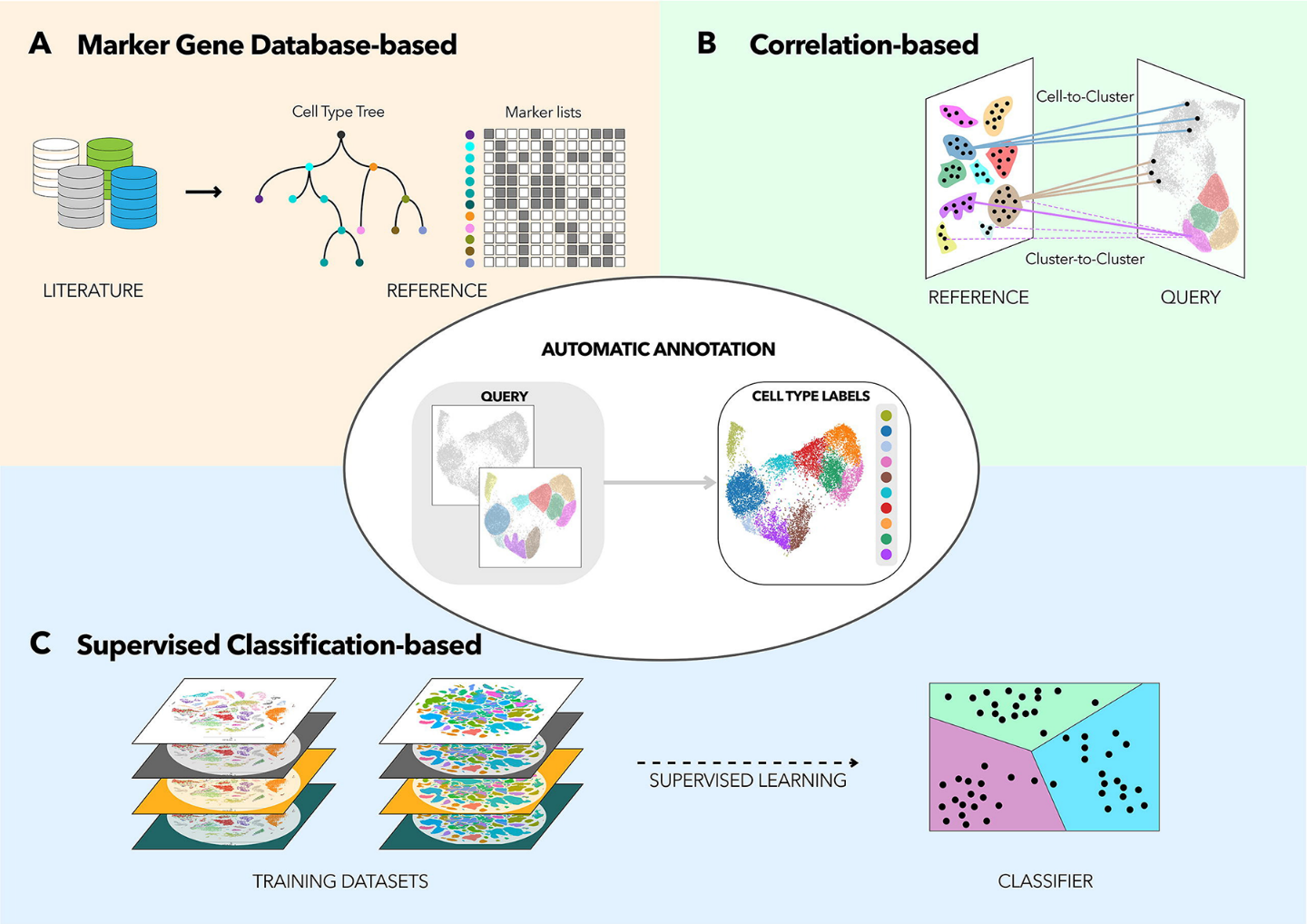


Figure 3.14: There are three main approaches to cell type annotation in single-cell studies, namely marker gene approaches, correlation approaches and supervised classification approaches. Image taken from Pasquini et al. <https://doi.org/10.1016/j.csbj.2021.01.015>

In all three approaches, the choice of reference data is crucial for the success of the cell type annotation.

Ideally, the reference data should encompass the cell types that are expected in the data to be annotated. For example, do not annotate a bone marrow dataset using a brain atlas reference.

The cell type labels in the reference data will determine the “resolution” of the predicted cell type annotation. Using a reference that only has coarse cell type labels (e.g.  a general T-cell label as opposed to CD4+ Naive, CD8+ Naïve, etc. labels) can only generate coarse annotations.

It is important to have the option of having unassigned cells. This happens when a cell (or cluster) has a low score across all the cell types in the reference. The cell could be of a novel, rare type that is not annotated in the reference data or could represent stressed or dying cells that have “lost” their cell identity.

Furthermore, a cell’s type can be plastic, and cells can trans-differentiate between two types. In this case, a cell (or cluster) might have high scores for two different cell types.

**3.4.2 Annotation via marker genes and functional analysis**

We will annotate the clusters we found previously based on marker genes and the HAY\_BONE\_MARROW gene signature overrepresentation analysis that we did earlier. The final annotations will be assigned manually.

The major cell states: HSC / progenitors / erythroid / dendritic cell / monocyte / B cell / T cell / NK cell can be determined from the HAY\_BONE\_MARROW signatures. Some cellular substates can be resolved using marker genes.

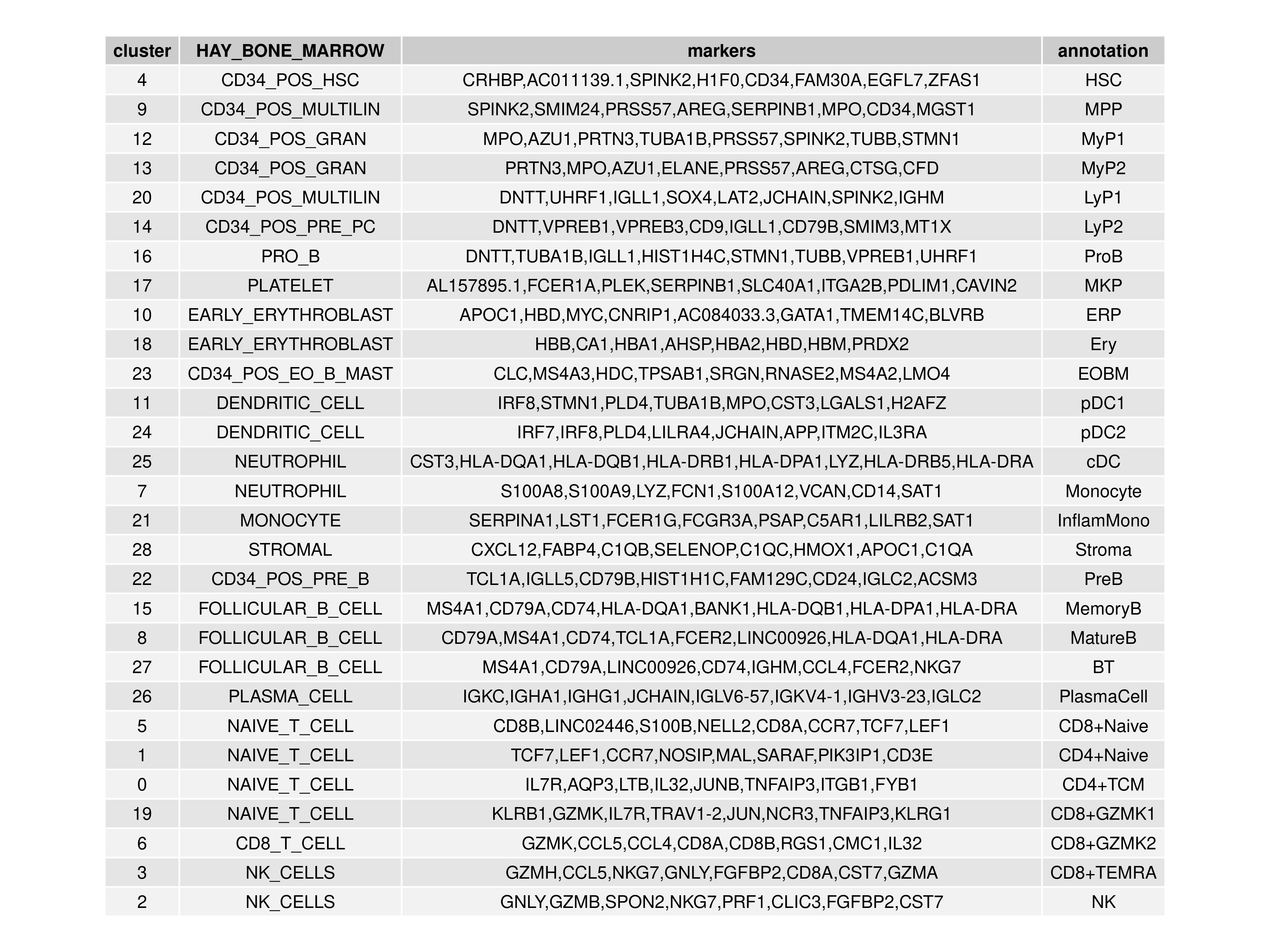


Figure 3.15: Annotation of the 29 clusters in the bone marrow dataset via marker genes and HAY\_BONE\_MARROW gene signature analysis.

**Code for the table above:**

This code extracts the most significant HAY\_BONE\_MARROW signature and top 8 marker genes by log-fold-change for each cluster and generates a table of annotations of the clusters.

library(gridExtra)

library(Seurat)

library(data.table)

oupAnnot = data.table(cluster = reorderCluster)

*# Add in top 5 marker genes*

ggData <- oupMarker[, head(.SD, 8), by = "cluster"]

ggData <- ggData[, paste0(gene, collapse = ","), by = "cluster"]

colnames(ggData)[2] <- "markers"

oupAnnot <- ggData[oupAnnot, on = "cluster"]

*# Add in the BONE\_MARROW enrichment from the results of enrichment*

*# analysis calculated previously (contained in oupMarkerFunc)*

ggData <- oupMarkerFunc[grep("BONE\_MARROW", ID)][, head(.SD, 1), by = "cluster"]

ggData <- ggData[, c("cluster", "ID")]

ggData$ID <- gsub("HAY\_BONE\_MARROW\_", "", ggData$ID)

colnames(ggData)[2] <- "HAY\_BONE\_MARROW"

oupAnnot <- ggData[oupAnnot, on = "cluster"]

*# Add in annotation*

oupAnnot$annotation <- c("HSC","MPP","MyP1","MyP2","LyP1","LyP2","ProB",

"ERP","Ery","MKP","EOBM",

"pDC1","pDC2","cDC","Monocyte","InflamMono","Stroma",

"PreB","MemoryB","MatureB","BT","PlasmaCell",

"CD8+Naive","CD4+Naive","CD4+TCM",

"CD8+GZMK1","CD8+GZMK2","CD8+TEMRA","NK")

*# Output table*

png("images/clustReAnnotTable.png",

width = 12, height = 9, units = "in", res = 300)

p1 <- tableGrob(oupAnnot, rows = NULL)

grid.arrange(p1)

dev.off()

saveRDS(seu, file = "bmSeu.rds")

**3.4.3 Visualizing annotated clusters**

The final step is to relabel the cell clusters into the annotation labels. The resulting annotated cell types can then be visualized in t-SNE and UMAP projections (Figure @ref{clust-annotTsum}).

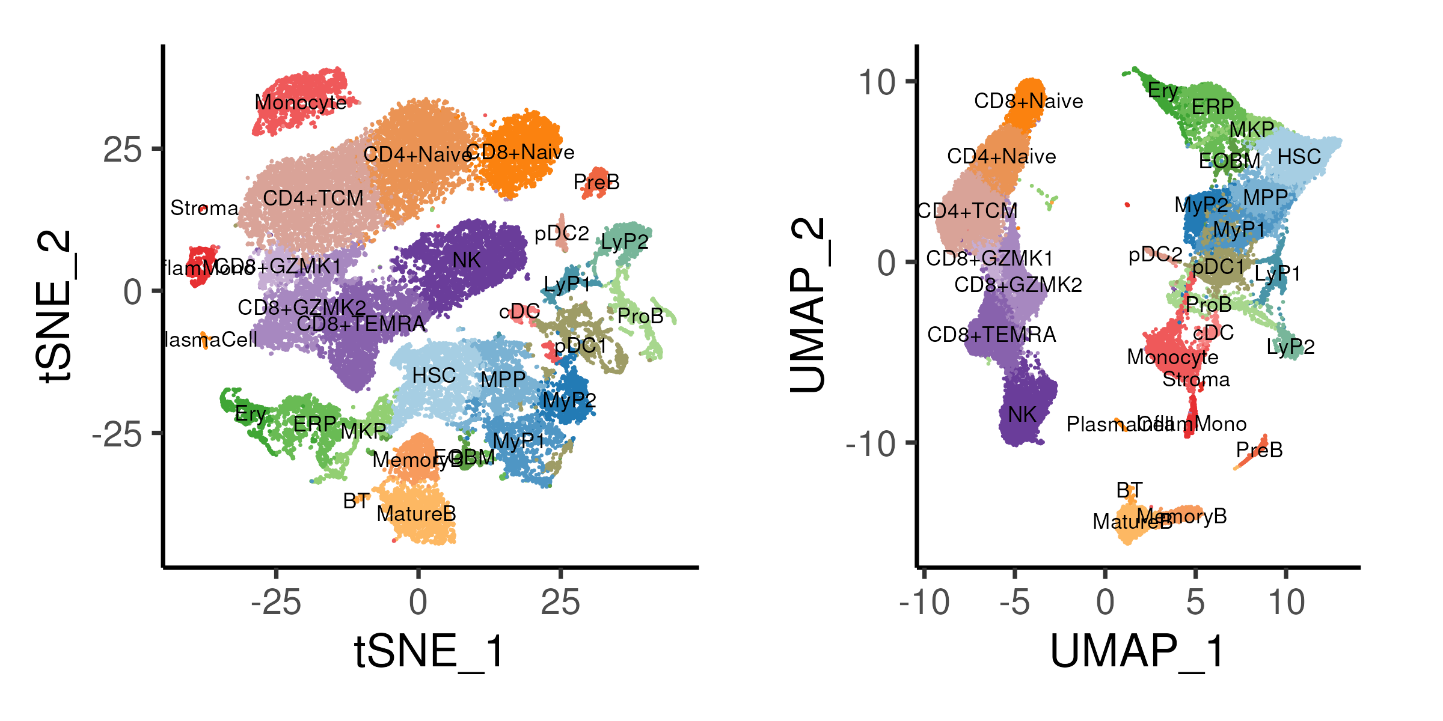


Figure 3.16: t-SNE and UMAP projections coloure by annotated cell types.

**Code for the plots above:**

We create a new metadata called celltype  in the Seurat object seu by mapping the cluster labels.

*# Add new annotation into seurat and replot*

tmpMap <- oupAnnot$annotation

names(tmpMap) <- oupAnnot$cluster

seu$celltype <- tmpMap[as.character(seu$cluster)]

seu$celltype <- factor(seu$celltype, levels = tmpMap)

Idents(seu) <- seu$celltype *# Set seurat to use celltype*

*# Plot ideal resolution on tSNE and UMAP*

p1 <- DimPlot(seu, reduction = "tsne", pt.size = 0.1, label = TRUE,

label.size = 3, cols = colCls) + plotTheme + coord\_fixed()

p2 <- DimPlot(seu, reduction = "umap", pt.size = 0.1, label = TRUE,

label.size = 3, cols = colCls) + plotTheme + coord\_fixed()

ggsave(p1 + p2 & theme(legend.position = "none"),

width = 8, height = 4, filename = "images/clustReAnnotTsUm.png")

*# Save Seurat Object at end of each section*

saveRDS(seu, file = "bmSeu.rds")